Universidade de Lisboa

Faculdade de Medicina de Lisboa

Co-transcriptional quality control of mRNA biogenesis: impact for human genetic diseases

Rita Catarina Vaz Drago

Orientador: Professora Doutora Maria do Carmo Salazar Velez Roque da Fonseca
Co-orientador: Prof. Doutora Noélia Maria Fernandes Custódio

Tese especialmente elaborada para a obtenção do grau de Doutor em Ciências Biomédicas
Especialidade de Biologia Celular e Molecular
2018

Universidade de Lisboa

Faculdade de Medicina de Lisboa

Co-transcriptional quality control of mRNA biogenesis: impact for human genetic diseases

Rita Catarina Vaz Drago

Orientador: Professora Doutora Maria do Carmo Salazar Velez Roque da Fonseca
Co-orientador: Prof. Doutora Noélia Maria Fernandes Custódio

Tese especialmente elaborada para a obtenção do grau de Doutor em Ciências Biomédicas
Especialidade de Biologia Celular e Molecular

Júri:

Presidente: Doutor José Luis Bliebernicht Ducla Soares, Professor Catedrático em regime de *tenure* e Vice-Presidente do Conselho Científico da Faculdade de Medicina de Lisboa.

Vogais:

- Doutor José Rueff Tavares, Professor Catedrático da NOVA *Medical School* – Faculdade de Ciências Médicas da Universidade Nova de Lisboa.
- Doutora Cecília Maria Arraiano, Investigadora Coordenadora, *Head of Controlo f Gene Expression laboratory* do Instituto de Tecnologia Química e Biológica António Xavier da Universidade Nova Lisboa.
- Doutora Paula Duque Magalhães Santos, Investigadora Principal, *Goup Leader*, *Plant Molecular Biology* do Instituto Gulbenkian de Ciência (IGC).
- Doutora Luísa Miranda Figueiredo, Investigadora, *Group Leader* no iMM João Lobo Antunes, Unidade de Investigação associada à FMUL.
- Doutora Maria do Carmo Salazar Velez Roque da Fonseca, Professora Catedrática da Faculdade de Medicina da Universidade de Lisboa – (*Orientadora*).
- Doutor João Pedro Taborda Barata, Professor Associado Convidado da Faculdade de Medicina da Universidade de Lisboa.

Instituição Financiadora: Fundação Para a Ciência e a Tecnologia (SFRH/BD/90231/2012).

2018

A Impressão desta tese foi aprovada pelo Conselho Científico da Faculdade de Medicina da Universidade de Lisboa em reunião de dia 19 de Abril de 2018

V - BRAÇO SEM CORPO BRANDINDO UM GLÁDIO

Entre a árvore e o vê-la
Onde está o sonho?
Que arco da ponte mais vela
Deus?... E eu fico tristonho
Por não saber se a curva da ponte
É a curva do horizonte...

Entre o que vive e a vida
Pra que lado corre o rio?
Árvore de folhas vestida —
Entre isso e Árvore há fio?
Pombas voando — o pombal
Está-lhes sempre à direita, ou é real?

Deus é um grande Intervalo,
Mas entre quê e quê?...
Entre o que digo e o que calo
Existo? Quem é que me vê?
Erro-me... E o pombal elevado
Está em torno na pomba, ou de lado?

Fernando Pessoa, «ALÉM-DEUS». *Orpheu*, nº 3. (Lisboa, 1916)

Preface

The present thesis entitled *Co-transcriptional quality control of mRNA biogenesis: impact for human genetic diseases* contains results of research work developed at Instituto de Medicina Molecular, Faculdade de Medicina de Lisboa, under the supervision of Professora Doutora Maria Carmo-Fonseca and Professora Doutora Noélia Custódio. This dissertation is divided in four chapters preceded by a Portuguese and an English resumes and keywords, and a list of abbreviations. Technical details, references and published articles are at the end of the thesis.

The first chapter corresponds to the general introduction of this dissertation, where I detailed the pathways involved in mRNA biogenesis, namely DNA transcription and pre-mRNA processing. I also described RNA quality control mechanisms that operate to degrade defective transcripts due to errors in pre-mRNA processing. Finally, I presented a brief overview of the technical methodologies to study mRNA biogenesis.

The objectives of this work were summarized in the second chapter.

The original data that resulted in this dissertation are presented in the third chapter that is divided in four main subchapters. In the first subchapter I explored the processes involved in mRNA quality control both in the nucleus and in the cytoplasm of patient-derived cells. In the second subchapter of results I reviewed and analysed the impact of deep-intronic mutations on human disease. In the third sub-chapter, I showed how different methodologies to isolate nascent transcripts may affect measurement of efficiency of splicing. Finally, using a single-molecule analysis methodology, I measured the time of release of transcripts from the site of transcription.

In the last chapter, a general discussion, focused on the implications of these discoveries for human disease, is presented.

Table of contents

Acknowledgments

Agradeço à Professora Carmo por ter aceitado receber-me no seu laboratório. Por me ter dado a oportunidade de assistir às suas aulas de Biologia Celular e aprender consigo a olhar para dentro de uma célula, no escuro e frio de uma sala de microscópios, e ver o mundo lá dentro.

Agradeço à Professora Noélia por tudo aquilo que me ensinou e inspirou a ser. Pela serenidade que sempre me transmitiu, pelo rigor com que pauta a sua carreira científica e pedagógica. Agradeço-te por tudo aquilo que a palavra amizade pode conter. À Sofia, Maitê e David por serem pequeninos e me ensinarem coisas enormes.

Aos membros do meu Comité de Tese, os Professores João Barata, Peter Jordan e Joana Desterro por me terem apoiado e guiado. Dirijo um agradecimento especial à Joana, por me ter ajudado tanto que estas palavras não chegam, porque és um exemplo e trazes sempre coisas boas à minha vida.

Dirijo uma palavra especial às Professoras Teresa Carvalho, Sandra Martins, Célia Levy, Evgenia Bekman, e aos Professores João Ferreira, Francisco Enguita e Sérgio de Almeida pelo tempo e ensinamentos que me concederam.

Ao Dr. Filipe, pela perseverança, boa-disposição e serenidade com que sempre temperou as nossas discussões.

Agradeço aos meus colegas Vanessa Pires, Tomás Gomes, Kenny Rebelo, Joana Tavares, Duarte Brandão, Soraia Silva, Marina Costa, Robert Martin, João Pessoa, Carla Gomes, Bruno Jesus, Simão Rocha, Ana Raposo, Maria Arez, Rui Luís, Teresa Silva, Marta Ribeiro, Pedro Barbosa pelas extensas discussões científicas e não-científicas que tivemos. Por todas as perguntas que me fizeram e por todas as respostas que me deram.

À Rita Mendes de Almeida, pelo teu sorriso aberto e por termos sido durante muito tempo só as duas. À Catarina Vale, que começou esta jornada comigo. Ao Pedro Prudêncio, com quem tanto aprendi.

À Filipa, por estar sentada na primeira fila daquele autocarro. Obrigada por me relembrares do que eu tendo a esquecer, por todos os momentos em que nos divertimos juntas.

À Sílvia, pela excepcional pessoa e cientista que é. Por estares sempre lá.

À Ana Jesus, pelo método e rigor com que sempre a vi trabalhar, pelas conversas sem princípio, fim ou definições, como *uma onda que se alevantou*.

Aos meus amigos, Pena, Idálio, João, Sara, Catarina Santos, pelos abraços e sorrisos.

À Ana Amaral, por despertar o meu melhor. À Filipa e ao Marco, pela amizade sem limites. Por terem sempre os braços abertos à chegada.

Aos meus colegas de há tanto tempo, Diana Lázaro, Soraia Rosa, Ana Calarrão, Nelson Afonso, Tiago Amado, Mara Nunes, Renata Castro, João Pereira, Sara Antunes, que partilharam comigo as primeiras lições.

Ao José Rino, António Temudo, Ana Nascimento, Patrícia Cúcio, Sérgio Marinho, José Albuquerque, Marisa Cabrita pela vossa sempre tão desinteressada ajuda. À Dinora por me conceder um cantinho no seu espaço.

Um agradecimento especial à Isa, Cynthia, Rui, Rayane e Arthur, pela amizade já longa e por se terem tornado na extensão da minha família. Estou-vos eternamente grata.

À minha família de anos, Graça, Neves, Vasco, D. Fernanda, Sr. Desidério, Fábio, Sr. Neves.

À Rosa, pela honestidade, delicadeza e suporte.

À Ana Paula, por ser parte de nós há tanto tempo que as fronteiras desapareceram. Obrigada por todos os abraços e teorias que inventámos.

À Tia Teresa, pela alegria contagiante e bondade com que sempre nos acolheu. À Tia Fernanda, por recordar o meu pai.

À minha avó Catarina, pela infindável capacidade de dar, por estar comigo apesar da distância que os anos impõem.

Ao meu avô Francisco, por me ter ensinado a fazer pastéis-de-nata com conchas e areia, pelos passeios de camião e pelos sorvetes no Verão. Por ter sido meu pai.

Ao meu irmão João, pela lucidez dos conselhos, pela paciência e riso fácil. Por me mostrar o valor da amabilidade, delicadeza e firmeza.

À minha irmã Mariana, por ter o sorriso mais bonito do mundo, dizendo que *o essencial é invisível aos olhos*.

Ao meu pai Afonso, por me ter dito que a Lua, em vez de crateras, tem uma face e sorri. Obrigada pelos serões demorados de histórias com nano-bichos, cavaleiros e estrelas distantes.

À minha mãe Isabel, por me ter oferecido o meu primeiro livro, uma estrela para levar ao peito e por me ouvir sempre. Obrigada por teres segurado no selim enquanto eu dava as primeiras pedaladas e pelos teus conselhos sempre ponderados e sábios. Por me ensinares a ser perseverante, forte, inquebrável.

Ao Sete.

Abbreviations

3' UTR	3' untranslated region
4sU	4-thiouridine
4tU	4-thiouracil
5' UTR	5' untranslated region
CFIm	cleavage factor I and II
CFIIm	cleavage factor II
CHX	cycloheximide
CPSF	cleavage and polyadenylation specificity factor
CstF	cleavage stimulation factor
CTD	carboxyl-terminal domain
EBV	Epstein-Barr virus
EJC	exon junction complex
DNA	deoxyribonucleic acid
GTF	general transcription factor
Biotin-HPDP	N-[6-(Biotinamido)hexyl]-3'-(2'-pyridyldithio)-propionamide
hTRAMP	human Trf4p/5p-Air1p/2p-Mtr4p polyadenylation
iPSC	induced pluripotent stem cells
LCL	lymphoblastoid cell line
miRNA	micro RNA
mRNA	messenger RNA
Biotin-MTS	2-((Biotinoyl)amino)ethylmethanethiosulfonate
NEXT	nuclear exosome targeting
NMD	nonsense mediated decay
RNA	ribonucleic acid
RNAPII	ribonucleic acid polymerase II
p(A)	polyadenylation
PABPN	nuclear poly(A) binding protein
PAP	poly(A)-polymerase
PAXT	poly(A) tail exosome targeting
PTC	premature termination codon
PyT	polypyrimidine tract
QC	quality control
RT-qPCR	reverse transcription quantitative real-time PCR
rRNA	ribosomal RNA
tRNA	transfer RNA
TS	transcription site

TSS	transcription start site
TTS	transcription termination site
siRNA	small interfering RNA
SM	splicing mutation
snoRNA	small nucleolar RNA
snRNA	small nuclear RNA
snRNP	small nuclear ribonucleoprotein
ss	splice site
SSA	spliceostatin A

Resumo

A expressão da informação codificante contida nas cadeias de DNA, implica a biogénese de moléculas de RNA mensageiro precursor (pré-mRNA) por um processo designado por transcrição, o processamento das referidas moléculas de pré-mRNA em moléculas de RNA mensageiro (mRNA), e culmina na síntese de proteínas por um processo designado por tradução. A compartimentalização celular dos eucariotas leva a que estes fenómenos ocorram em locais específicos na célula: a transcrição e o processamento ocorrem no núcleo, ao passo que a tradução ocorre no citoplasma, o que implica e transporte núcleo-citoplasmático das moléculas de mRNA. Transcrição e processamento envolvem a actividade específica de diferentes maquinarias celulares que comunicam entre si de forma a coordenar a biogénese das moléculas de mRNA.

A transcrição de genes que codificam proteínas é levada a cabo pela RNA Polimerase II (RNAPII) e consiste num processo sequencial que inclui iniciação, elongação e terminação das cadeias de RNA nascente. Durante a transcrição, o pré-mRNA sofre três tipos de processamento: *capping*, que consiste na modificação química da extremidade 5'; *splicing*, que compreende a redução do comprimento das cadeia de RNA nascente através da clivagem de sequências intercalares (intrões) e junção das sequências codificantes (exões); *cleavage*, poliadenilação e *release*, colectivamente designados por *3' end processing*, que consiste na clivagem do RNA nascente, adição de uma cauda de Adeninas [poli(A)] na extremidade 3' e libertação da cadeia de DNA molde. Após a ocorrência do *release*, as moléculas de mRNA difundem no nucleoplasma até encontrarem um poro nuclear e são exportados para o citoplasma.

Ao contrário do mecanismo de *capping*, as reacções de *splicing* e *3' end processing* dependem do reconhecimento de sequências conservadas presentes no pré-mRNA por parte de complexos proteicos especializados. No mecanismo de *splicing*, essas sequências estão localizadas nas fronteiras entre os exões e intrões e são designadas por *splice sites*, já no caso do *3' end processing*, essas sequências estão localizadas na região 3' do transcrito. Mutações nessas sequências ou nos genes que codificam os componentes dos

complexos que catalisam as reacções de processamento podem comprometer a eficiência das referidas reacções, resultando na produção de moléculas de mRNA defeituosas. Se traduzidos, os transcritos mutantes poderiam dar origem a proteínas com efeito potencialmente deletério para a célula. No entanto, isso raramente ocorre porque as células eucariótas desenvolveram mecanismos de controlo da qualidade de moléculas de mRNA, prevenindo que moléculas defeituosas sejam traduzidas. Mutações que afectam o *splicing* resultam geralmente na introdução de codões de terminação da tradução prematuros (*premature termination codons* - PTCs) no mRNA. Moléculas de mRNA que contêm PTCs são reconhecidas e degradadas por uma via citoplasmática dependente de tradução, designada por *nonsense mediated decay* (NMD). No núcleo de células de mamífero foram também identificadas vias de degradação nucleoplasmática de RNAs mutantes, e ainda outros mecanismos de controlo de qualidade que previnem a libertação de transcritos com erros no processamento do respectivo DNA molde e outros que bloqueiam a exportação de transcritos mutantes.

Estima-se que 30% de todas as doenças genéticas humanas sejam causadas por mutações que perturbam o mecanismo de *splicing*, no entanto, à excepção da via de degradação por NMD, pouco se sabe sobre mecanismos de controlo de qualidade de mutantes de *splicing* no contexto de doença genética.

As vantagens do controlo de qualidade do mRNA começaram a ser apreciadas na β-talassemia, pois foi constatado que, na maioria dos casos, apenas os sujeitos homozigóticos sofriam de anemia grave. Por outro lado, os heterozigóticos tendiam a ser fenotipicamente saudáveis porque a via NMD previne a produção de formas truncadas de β-globina. No entanto, algumas mutações conferem um fenótipo dominante quando a via de NMD é contornada, indicando o seu papel crítico na neutralização de potenciais mutações com ganho de função. Curiosamente, foi também num modelo celular de β-talassemia que foi descrito, pela primeira vez, um mecanismo de controlo de qualidade que opera no núcleo promovendo a retenção de transcritos de β-globina com defeitos no processamento junto ao local de transcrição.

Com base nesta observação, coloquei a hipótese de que um ponto de verificação da qualidade do mRNA opera no núcleo, antes do NMD, para

reduzir a acumulação de transcritos mutados e assim impedir a produção de proteínas deletérias causadora de doença. De acordo com essa hipótese, transcritos com erros no processamento gerados por mutações em *splice sites* não são libertados da cromatina, podendo subsequentemente retroalimentar negativamente a transcrição e reduzir assim a síntese de moléculas de mRNA mutadas.

Como modelo de estudo, usei linhas celulares linfoblastóides derivadas de pacientes (hemi ou homozigóticos) com doenças genéticas causadas por mutações em *splice sites* canónicos e mutações na região codificante que introduzem um PTC. A introdução de PTC é frequente nos transcritos com mutações em *splice sites* como consequência da alteração do padrão de *splicing* por *exon skipping*, *intron retention* ou activação de um *splice site* críptico próximo.

As doenças genéticas estudadas foram: síndrome de Barth, surdez autossómica recessiva 49 e Xeroderma pigmentosum. No total, analisei por RT-qPCR a quantidade de RNA citoplasmático, nucleoplasmático e associado à cromatina de oito linhas celulares derivadas de pacientes e de uma derivada de um dador saudável. Para cinco das seis linhas com mutações em *splice sites* observei uma quantidade de RNA significativamente reduzida na fracção nucleoplasmática, resultado que sugere que transcritos mutantes são degradados no núcleo de células derivadas e pacientes com doenças genéticas. Adicionalmente, em três das linhas linfoblastóides encontrei níveis reduzidos de RNAs mutantes associados à cromatina. Com o objectivo de determinar se a redução na quantidade de RNA nascente se correlacionava com um decréscimo da actividade transcricional, optimizei um ensaio para medir a actividade transcricional com base no fornecimento de um análogo de uridina, *4-thiouridine* (4sU), às células em cultura. Este nucleótido modificado é incorporado nas moléculas de RNA nascente durante breves minutos, o que permite purificar esta subpopulação de transcritos, medindo assim actividade transcricional dos genes de interesse. Após quantificação por RT-qPCR, observei que a menor abundância de RNA associado à cromatina se correlacionava com a redução da actividade transcricional em duas das linhas analisadas. Ao bloquear a via de NMD com fármacos comprovei ainda que a redução dos níveis celulares de transcritos sintetizados a partir de genes com

mutações em *splice sites* é um mecanismo independente do NMD. Concluí assim que um mecanismo de controlo de qualidade do mRNA está activo no núcleo e leva à redução da transcrição de genes com mutações localizadas em *splice sites* canónicos.

As variantes que causam doença genética por alterarem o padrão de *splicing* estão maioritariamente localizadas em *splice sites* canónicos, no entanto a sequenciação das regiões intrónicas dos genes revelou que as mutações localizadas a mais de 100 bp de um exão, denominadas de *deep-intronic* podem ser também a causa de doenças genéticas humanas. Com o objectivo de avaliar o impacto das mutações *deep-intronic* na biogénese de moléculas de mRNA, fiz uma revisão da literatura onde são apresentados dados de RT-qPCR de mRNA extraído de células linfoblastóides e dados de sequenciação completa do genoma de pacientes com mutações *deep-intronic*.

Verifiquei que as variantes *deep-intronic* patogénicas são a causa de mais de 75 doenças monogénicas, bem como de síndromes hereditárias de cancro. Enquanto que as mutações em *splice sites* canónicos levam ao não reconhecimento daquele sinal, mutações localizadas no interior de intrões podem levar à criação e posterior reconhecimento de novos *splice sites* intrónicos, denominados de não-canónicos, em aproximadamente 70% dos casos. A criação, por mutação, de um *splice site* não-canónico, leva à activação de um *splice site* adjacente e à subsequentemente inclusão de uma sequência intrónica no mRNA designada por pseudo-exão. A mutação de elementos *intrónicos* que são reconhecidos por proteínas reguladoras de *splicing* também pode levar à inclusão de pseudo-exões no mRNA. Além disso, mutações *deep-intronic* podem afectar motivos reguladores da transcrição e genes de RNA não codificantes.

Uma vez que erros no mecanismo de *splicing* são a causa de uma elevada proporção de doenças genéticas, a medição da eficiência de *splicing* deve ser feita com grande precisão de forma a ter uma correcta compreensão dos mecanismos moleculares que podem estar afectados no contexto de doença.

Uma variedade de abordagens tem sido usada para purificar moléculas de RNA recém-transcritas e determinar a eficiência de *splicing*. Estudos prévios sugerem que a maior parte dos transcritos sofre *splicing* enquanto a RNAPII

transcreve o respectivo gene. Assim, a purificação de transcritos recém-sintetizados pode aumentar em muito a precisão da determinação da eficiência de *splicing*.

A marcação metabólica de RNA utilizando 4sU tem sido utilizada para purificar moléculas recém-transcritas e determinar a sua cinética de processamento. Esta abordagem baseia-se no tratamento com um reagente *thio*-reactivo para biotinilar o RNA marcado, que é então purificado por afinidade com estreptavidina. No entanto, a reacção de 4sU com o reagente comumente utilizado, designado por biotina-HPDP, é pouco eficiente, o que pode levar a que RNAs mais longos sejam purificadas em detrimento de RNAs mais curtos com menos número de moléculas de 4sU incorporadas. Com base neste resultado coloquei a hipótese de que a cinética de remoção de intrões longos é selectivamente subestimada em estudos em que o reagente biotina-HPDP é utilizado. Com o objectivo de testar esta hipótese, utilizei uma estratégia de biotinilação mais eficiente que usa o reagente *thio*-reactivo MTS. Mostrei que o RNA nascente purificado com biotina-HPDP contém uma proporção significativamente maior de intrões longos não processados em comparação com RNAs purificados com biotina-MTS, argumentando favoravelmente em relação à hipótese colocada. Estes resultados indicam ainda que a marcação com 4sU não interfere com a eficiência de *splicing* e que intrões humanos que variam em tamanho entre 240 e 13000 nucleótidos são *spliced* com uma eficiência idêntica.

A disrupção do *3' end processing* está também associada a um conjunto de doenças humanas. Contudo, comparado com o *splicing*, esta etapa da biogénese do mRNA tem sido menos estudada. Com o objectivo de compreender com maior detalhe este mecanismo de processamento, decidi estudar a cinética de libertação da cromatina de moléculas de RNA de β-globina e IgM utilizando microscopia confocal. Tirando partido de dois métodos de marcação de moléculas de RNA que consistem na inserção de locais de ligação de proteínas fágicas (MS2 ou PP7) a montante do local de poliadenilação, foi possível realizar um estudo em células vivas e avaliar o comportamento de moléculas individuais de RNA. Mostrei que a libertação do local de transcrição dos RNAs nascentes ocorre em cerca de 15-30 segundos, tempo necessário para clivar e libertar o RNA nascente totalmente transcrito do

local de transcrição. Como controlo, determinei que a diminuição dos níveis celulares do factor de clivagem de pré-mRNA CPSF3 se reflecte no aumento do tempo de permanência no local de transcrição de ambos os transcritos. Utilizando um método de marcação de RNA diferente (λN22) em que os locais de ligação da proteína fágica foram inseridos a jusante do local de poliadenilação do mini-gene IgM, medi também que o tempo de terminação da transcrição varia entre 20 e 80 segundos.

Este estudo permitiu determinar pela primeira vez a cinética de *release* de transcritos-repórter e terminação da transcrição, para moléculas individuais e em células vivas. Esses resultados têm importantes implicações para uma compreensão mecanicista da biogénese do mRNA.

No seu todo, os dados originais que resultaram nesta dissertação detalharam os processos envolvidos na síntese e decaimento das moléculas de RNA nos contextos de saúde e doença.

Palavras-chave

Transcrição

Splicing

3' end processing

Controlo de qualidade co-transcricional

Doença genética

Summary

Protein coding genes are transcribed in the nucleus by RNA polymerase II (RNAPII) forming a precursor messenger RNA (pre-mRNA) that undergoes extensive processing including 5' capping, splicing, 3' end cleavage and polyadenylation to form a mature mRNA. Pre-mRNA processing takes place co-transcriptionally, potentiated by the carboxyl-terminal domain (CTD) of the largest subunit of RNAPII, in a way that transcription and processing machineries communicate with each other to coordinate mRNA biogenesis. After being released from the chromatin template, mRNAs diffuse through the nucleoplasm until they encounter a nuclear pore to be translocated to the cytoplasm where they are translated into proteins, the final outcome of gene expression.

Mutations that alter the coding sequence or affect splicing often result in the introduction of premature termination codons (PTCs). If translated, the resulting mRNAs would give rise to truncated proteins with potential deleterious effect for the cell. However, this rarely occurs because eukaryotic cells are able to recognize and degrade mRNAs containing PTCs by a cytoplasmic pathway referred to as nonsense-mediated mRNA decay (NMD). NMD was the first reported example of a quality control mechanism of gene expression. The advantages of mRNA quality control started to be appreciated in the case of beta-thalassemia, as it was found that in most cases only homozygotes suffered from severe anemia. Heterozygotes tend to be phenotypically healthy because NMD prevents production of truncated forms of beta-globin. In addition, thalassemia-like beta-globin mutations resulting in mRNA processing defects induce a nuclear RNA surveillance mechanism that lead to the retention of RNAs near the transcription site. To study the quality control mechanisms that operate during mRNA biogenesis it is essential to fully understand the process of gene expression in health and disease. One pertinent question that was addressed in my PhD work was how general the co-transcriptional mRNA quality control mechanism is and what is its impact in human genetic diseases.

To address this question, I used as model system lymphoblastoid cell lines from patients with genetic diseases caused by splicing mutations and

mutations in the coding region that introduce a PTC. Quantification of nascent transcripts revealed that a subset of genes containing splicing mutations have reduced transcriptional activity. Inhibition of NMD did not alter the levels of chromatin-associated transcripts, suggesting that a transcription-coupled surveillance mechanism operates independently from NMD to reduce cellular levels of abnormal RNAs in the context of human genetic diseases.

Disease-causing mutations that disrupt splicing are mostly localized in splice sites, however next-generation sequencing has revealed that mutations localized deep within introns (more than 100 base pairs away from exon-intron junctions) can be the cause of human genetic diseases. Aiming to highlight the importance of studying variation in deep intronic sequences, I reviewed evidence from mRNA analysis and entire genomic sequencing indicating that deep-intronic pathogenic mutations are the cause of over 75 monogenic disorders as well as hereditary cancer syndromes. Interestingly, deep-intronic mutations most commonly create/activate non-canonical splice sites in the pre-mRNA molecule that subsequently lead to pseudo-exon inclusion in the mature mRNA.

Since disruption of splicing causes approximately 30% of human genetic diseases, measurement of splicing efficiency is essential to understanding gene regulation in wild-type and splicing-mutated genes. A variety of approaches have been used to purify nascent transcripts and determine the efficiency of splicing. Specifically, purification of newly transcribed molecules using 4sU-tagging has been widely used. Classically, this approach relies on treatment with a thio-reactive reagent HPDP to biotinylate the tagged RNA, which is then affinity-purified with streptavidin. Taking advantage of an efficient biotinylation strategy that uses MTS reagent, I showed that nascent RNA purified with biotin-HPDP contains a significantly higher proportion of unspliced long introns compared to RNAs purified with MTS-biotin. This argues that the splicing kinetics of long introns may be selectively under-estimated in studies using biotin-HPDP, which may lead to mis-calculation of processing efficiency in different biological contexts.

Disruption of 3' end processing can also be the cause of many human disorders. However, compared to splicing, this step of mRNA biogenesis has been less studied. To further study 3' end processing and transcription

termination, I used a live-cell and single-molecule approach, in which time of release of two different reporter transcripts from the transcription site (TS) was measured. By using two different RNA labelling methods, MS2 and PP7, I showed that β-globin and IgM transcripts are released within 15-25 seconds after transcription of the 3' end of the gene. Furthermore, I showed that downregulating the cleavage factor CPSF3 by RNAi increases time of permanence at TS of both transcripts. Using a different RNA labelling method inserted past the poly(A) site (λN22), I determined that the time of transcription termination ranges between 20-80 seconds, with an average of 30 seconds. These results have important implications for a mechanistic understanding of mRNA biogenesis, particularly at 3' end.

Altogether, the original data that resulted in this dissertation detailed the processes involved in mRNA synthesis and decay in the contexts of health and disease.

Keywords

Transcription

Splicing

3' end processing

Co-transcriptional quality control

Genetic disease

1. Introduction

1.1. The human genome and transcriptome

The human genome encodes approximately 58,000 genes that precisely regulate all cellular functions (Consortium 2012). Among those, protein-coding genes constitute the more representative class of human genes, accounting for 19,000 and spanning around 30% of the human genome (Consortium 2012). On average, such class of genes are composed of 8-10 exons interrupted by much larger non-coding sequences called introns. In fact, exons are the only sequences that will be converted in an amino-acid sequence and correspond to approximately 2% of the human genome.

In addition to protein-coding genes, the human genome also encodes different classes of non-coding genes, these include long non-coding genes that are genes with a structure similar to protein-coding genes but do not code for proteins, micro RNA (miRNA) genes and short interfering (siRNA) genes that are involved in gene silencing, ribosomal RNA (rRNA) genes and transfer RNA (tRNA) genes that are directly involved in the production of proteins, small nucleolar RNA (snoRNA) genes that drive chemical modification in snRNA, rRNA and tRNA, and small nuclear RNA (snRNA) genes that are involved in protein-coding RNA molecules.

Through a process called transcription, the genes encoded in the human genome give rise to RNA molecules. Protein-coding genes and some non-coding genes are transcribed by RNA polymerase II (RNAPII). The remaining genes are transcribed by RNA polymerase I (RNAPI) and RNA polymerase III (RNAPIII). RNAPI produces rRNA precursor for the mature 25/28S, 18S and 5.8S rRNAs, whereas RNAPIII synthesizes small structured RNAs like tRNA, spliceosomal U6 small nuclear RNA (snRNA), ribosomal 5S rRNA and 7 SL RNA (Griesenbeck, Tschochner et al. 2017). Thus, in mammalian cells all RNA polymerases are involved in the production of non-coding RNAs. Transcriptomic analysis of different types of cells reveals that 90% of the total cellular RNA consists of rRNA and tRNA, while the RNA that codes for protein only represents 1-3% of the total cellular RNA content (Palazzo and Lee 2015).

The set of RNA molecules produced in a certain cell or organism constitute the transcriptome. The composition of the transcriptome actively and dramatically changes, depending on the state of development, environmental and disease conditions or drug treatments

DNA and RNA molecules share many features but there are many others that are specific of each group of molecules. Both are composed of nitrogenous bases covalently bound to a sugar-phosphate backbone but the sugar in RNA is ribose and in DNA is deoxyribose. RNA and DNA share three of the four bases that compose nucleic acids (adenine, cytosine and guanine), the fourth base thymine and uracil are specifically added to DNA and RNA molecules, respectively. Double-stranded genomic DNA molecules are arranged in a polymeric complex called chromatin. The fundamental unit of chromatin is the nucleosome, which repeats every 160 to 240 bp across the genome. Each nucleosome contains a nucleosome core, composed of an octameric complex of the core histone proteins, which forms a spool to wrap 145 to 147 bp of DNA [reviewed in (Richmond and Davey 2003, McGinty and Tan 2015)]. Although RNAs are usually single-stranded molecules that can associate with a wide range of RNA-binding proteins, they can also form double-stranded structures with other molecules or within the same molecule which gives rise to secondary structures within the RNA molecule. The secondary structures are of high importance to the cellular roles different classes of RNAs play in the cell. Additionally, RNA molecules, are susceptible to different and multiple kinds of chemical modifications and can travel between the nucleus and the cytoplasm.

1.2. Biogenesis of messenger RNA

To produce a protein, a protein-coding gene has to be transcribed into an RNA molecule that is subsequently translated into a protein. In prokaryotes, these processes are spatially and temporally coupled. In eukaryotes, however, the presence of a nuclear envelope physically separates transcription that occurs in the nucleus from translation that takes place in the cytoplasm. Additionally, eukaryotic RNA molecules undergo several processing steps from their synthesis until they are ready for translation, making the translatable molecules dramatically different from the products of transcription: they are much shorter and chemically modified at 5' and 3' ends. Thus, biogenesis of a mature transcript, called messenger RNA (mRNA), comprises synthesis and processing of a precursor mRNAs (pre-mRNA). The resultant mature mRNA has a tripartite structure consisting of a 5' untranslated region (5' UTR), a coding region made up of triplet codons that each encode an amino acid and a 3' untranslated region (3' UTR).

Synthesis and processing of pre-mRNA molecules are performed by distinct and specialized cellular machineries that interact in time and space in a tightly regulated way.

As previously mentioned, transcription of a protein-coding gene is performed by the 12-subunit protein complex RNAPII and includes initiation, elongation and termination of transcription. A protein-coding gene is transcribed from the transcription start site (TSS) that marks the 5' end of the first exon until the polyadenylation (polyA or pA) site that marks the 3' end of the last exon. After reaching the 3' end of the gene, RNAPII can leave the DNA template or bind again to the promoter region to restart transcription (gene-looping model). In parallel, processing reactions are simultaneously carried out by a variety of processing factors that act on the nascent transcripts as RNAPII proceeds through the gene body and include 5' end processing, splicing and 3' end processing (**Figure 1**).

The kinetic coupling between these mechanisms is carried out by the C-terminal domain (CTD) of the largest core subunit of RNAPII. In humans, the

CTD of RNAPII is an unstructured and evolutionary conserved domain that comprises 52 tandem repeats of the consensus heptapeptide YSPTSPS. As RNAPII reads the DNA template, the CTD is subject to extensive phosphorylation at different amino acid residues, resulting in a phosphorylation code that contributes to regulation of RNAPII function as well as the recruitment of regulatory and processing factors at the right place and time (Harlen 2016).

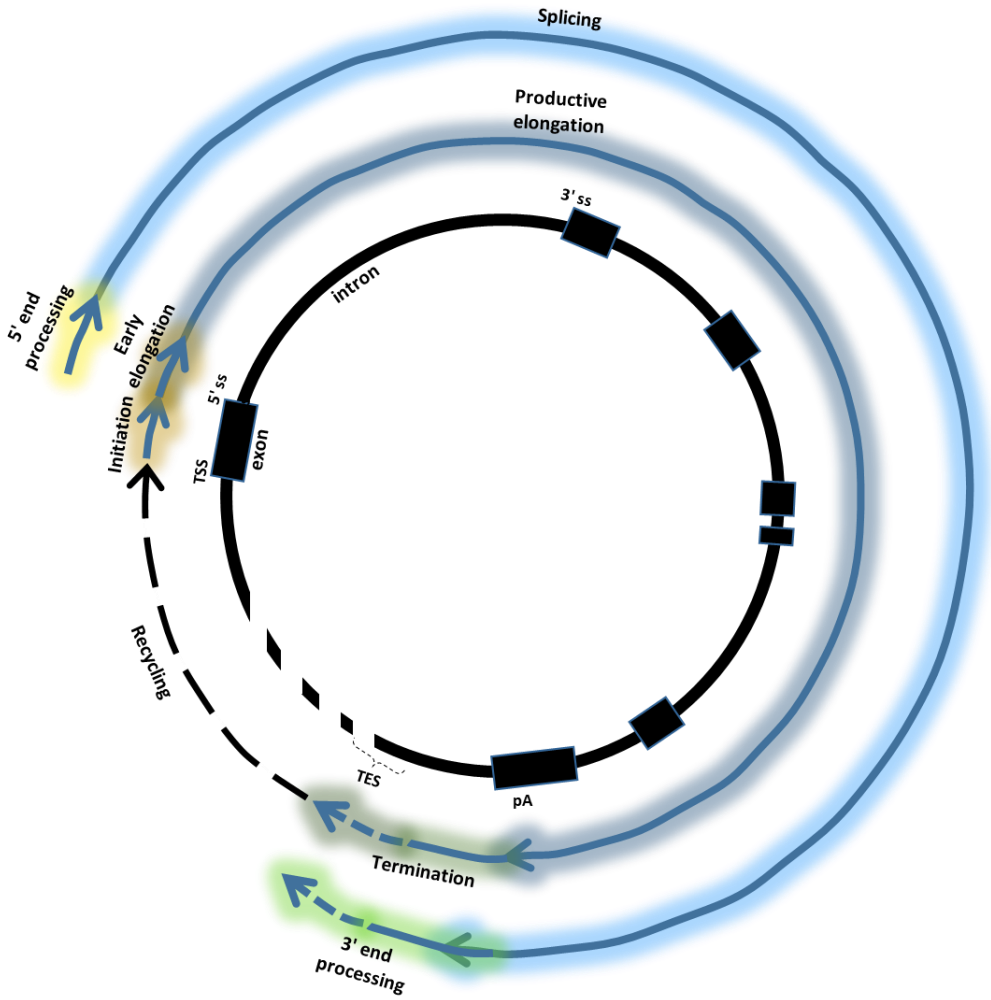


Figure 1 - Schematic view of coupling between RNAPII transcription and pre-mRNA processing.

A representative gene structure is shown in the inner circumference in which exons are represented by black boxes and intron by lines. The steps of the transcription cycle, from initiation to termination, are shown in the middle circumference. The dashed blue arrow indicates termination of transcription and dissociation of RNAPII. The dashed black arrow represents the recycling of RNAPII, reinitiating transcription on the same DNA (gene looping). Mechanisms involved in pre-mRNA processing are shown in the

outer circumference. TSS – transcription start site, TES – transcription end site, ss – splice site, pA – polyadenylation site

Initiation of transcription, early elongation and 5' end processing

Expression of protein-coding genes is controlled by RNAPII, however the polymerase does not initiate promoter-specific transcription on its own. Rather, RNAPII initiation is regulated by a protein macromolecular complex known as the pre-initiation complex (PIC), which is required for the opening of the duplex DNA and identification of the start site for transcription and consists of RNAPII itself, six General Transcription Factors (GTFs) and the Mediator (Bernecky, Herzog et al. 2016, Cramer 2016, He, Yan et al. 2016, Nogales, Louder et al. 2017).

Each of the six GTFs (TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH) have varied functional and structural roles. For example, TFIID is the first GTF recruited to the promoter region and contains the TATA-binding protein (TBP) that recognizes the initiator element. TFIIH also plays important roles in initiation of transcription by RNAPII: it triggers melting of the DNA molecule and adds a phosphate group to the serine 5 residue of the RNAPII-CTD, which is read as a signal to start polymerization of pre-mRNA [reviewed in (Han and He 2016)].

The Mediator is composed of 26 subunits in humans, although subunits can be lost or added depending on its biological function [reviewed in (Allen and Taatjes 2015)]. Mediator functions as a bridge between RNAPII and GTFs, since it communicates regulatory signals from DNA-bound GTFs directly to the CTD of RNAPII. Additionally to regulation of transcription initiation, Mediator has the ability to control pausing, elongation and recycling of RNAPII and also chromatin architecture organization [reviewed in (Poss, Ebmeier et al. 2013)]. Although Mediator exists in all eukaryotes, a variety of Mediator functions seem to be specific to metazoans, which is indicative of more diverse regulatory requirements (Sainsbury, Niesser et al. 2013, Allen and Taatjes 2015).

During transcription initiation, a promoter checkpoint mechanism that relies on the binding of TFIIB can trigger RNAPII to terminate prematurely (Liu,

Bushnell et al. 2011, Nechaev and Adelman 2011). When the nascent transcript is more than 10 nucleotides, TFIIB is ejected and RNAPII is stably engaged in transcription. At this point, RNAPII undergoes promoter escape by losing contact with the GTFs and enters elongation. Transcription elongation is composed of two distinct stages: early and productive elongation. Early elongation is defined as the transition between clearance of RNAPII from the promoter-bound GTFs and pausing of RNAPII after it transcribes around 25-50 nt. This pausing is facilitated by the DRB sensitivity-inducing factor (DSIF) and negative elongation factor (NELF) protein complexes and by the first nucleosome. While RNAPII is paused in the promoter region, two important reactions can take place at protein and RNA levels: phosphorylation of the pausing factors DSIF and NELF and of the CTD at Ser5 residue, and 5' end processing of the pre-mRNA. These reactions promote escape of RNAPII into productive elongation and are mediated by the heterodimer complex P-TEFb that can be recruited to paused polymerase by other protein complexes, such as an activator bound to a specific sequence in the DNA, BRD4 that recognizes and binds to acetylated histone tails and, as previously mentioned, by the Mediator (Kwak and Lis 2013, Jonkers, Kwak et al. 2014).

Processing of the 5' end of the pre-mRNA molecule is the first step in the biochemical processing of nascent pre-mRNAs, occurs as soon as RNAPII transcribes 20-100 nucleotides, coincides with promoter-proximal pause and involves the mRNA-capping enzyme that cleaves the 5' triphosphate of the first nucleotide of pre-mRNA and adds a guanosine monophosphate that is subsequently methylated by the guanine-7-methyltransferase (Mullen and Price 2017). The chemically modified 5' end of the pre-mRNA is than recognized and bound to the capping binding complex (CBC), composed CBP80 and CBP20 proteins, that can also interact with P-TEFb [reviewed in (Bentley 2014)].

A balance between NELF, DSIF and chromatin structure as well as factors that promote release of paused RNAPII by recruiting P-TEFb may determine the level of pausing and productively elongating RNAPII. The term “pausing” results from the interpretation of data from RNAPII immunoprecipitation associated with the chromatin and labelling of nascent RNA assays that reveal a promoter-proximal enrichment of RNAPII (Danko, Hah et al. 2013). However,

variation in RNAPII density depends on multiple parameters: elongation rate (bases added per unit of time), initiation frequency (number of start events per unit of time that result in productive elongation), and processivity (fraction of polymerases remaining on the template after each catalytic event) (Ehrensberger, Kelly et al. 2013). Single-molecule footprinting studies, suggested a model where absence of elongation is due to a continuous wave of initiation characterized by premature termination and rapid turnover of RNAPII instead of stalling of a pool of RNAPII. Consequently, transition to elongation would not be regulated by the release of a pool of engaged/paused RNAPII, but it would be mediated by the redirection of moving RNAPII molecules (Krebs, Imanci et al. 2017).

Elongation of transcription and splicing

After RNAPII is released from the promoter-proximal pause site, it starts productive elongation. Transcription elongation is highly variable as it is influenced by many physical and chemical environments along the gene body. The presence of nucleosomes, certain histone marks and G-rich DNA sequences can influence the rate and processivity of RNAPII during productive elongation (Venkatesh and Workman 2015). Histone chaperones and nucleosome remodelers play an important role in facilitating RNAPII movement through the gene body. Interestingly, RNAPII rate can be also negatively influenced by the presence of exons that show increased nucleosome-occupancy levels with respect to introns (Tilgner, Nikolaou et al. 2009).

The RNA that is tethered to DNA by an elongating RNAPII frequently undergoes splicing, a process that is responsible for shortening of a precursor mRNA molecule by removing the non-coding pieces of the precursor transcript (introns) and binding together the coding regions (exons) (Gilbert 1978).

The discovery that genes are arranged in pieces established an additional step in the pathway of production of a protein (Chow, Gelinas et al. 1977). To form a contiguous coding sequence that can be translated into a protein, introns have to be precisely removed from the pre-mRNA.

Excision of introns is catalyzed by a highly sophisticated ribonucleoprotein machinery called the spliceosome (Papasaikas and Valcarcel 2016). Intron-exon boundaries are delimited by short consensus sequences at the 5' (donor) and 3' (acceptor) splice sites (ss) that mediate recognition by the spliceosome; in addition, spliceosomal components interact with a catalytic adenosine (the branch point) and a polypyrimidine tract (PyT) located between the branch point adenosine and the 3' ss (**Figure 2**).

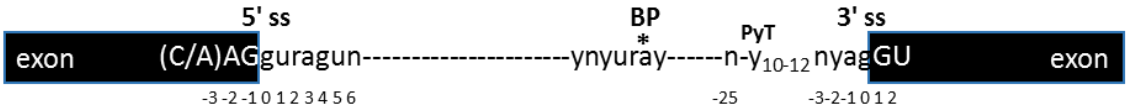


Figure 2 - Human consensus splice site sequences.
The most conserved nucleotide sequence is indicated for canonical 5' (donor) and 3' (acceptor) splice sites (ss), the branch point (a*) and polypyrimidine tract (PyT). The donor site (5' ss) is defined by the three terminal nucleotides of each exon and the first seven bases of the downstream intron. The acceptor site (3' ss) consists of the first two bases of the exon and the last 26 bases of the upstream intron, including the PyT. The coordinate ranges for the donor and acceptor site positions are [-3, +6] and [-25, +2], with a zero coordinate at the first intronic base of each splice site (Caminsky, Mucaki et al. 2014).

The spliceosome is formed by five small RNAs (the U1, U2, U4, U5, and U6 snRNAs) and more than 200 proteins (Wahl, Will et al. 2009). A subset of these proteins associates with the snRNAs forming functional particles called the U1, U2, U4, U5, and U6 snRNPs. Each snRNP consists of a snRNA molecule associated with a common set of Sm proteins and a variable number of additional specific proteins. Within the spliceosome, the consensus splicing sequences in the pre-mRNA are forced into 3-dimensional arrangements that enable the activation of an RNA catalytic center and trigger the splicing reaction (Hang, Wan et al. 2015).

Spliceosome assembly starts with the recognition of the 5' ss by the U1 snRNP (Seraphin and Rosbash 1989). Subsequently, the U2 snRNP associates with the branch point region. However, this interaction requires prior binding of splicing factor 1 (SF1) to the branch point sequence (Berglund, Chua et al. 1997, Liu, Luyten et al. 2001) and recruitment of U2 auxiliary factor U2AF to the 3' ss (Zamore, Patton et al. 1992). U2AF is a heterodimer composed of a 65

kDa subunit (U2AF65) that contacts the polypyrimidine tract and a 35 kDa subunit (U2AF35) that recognizes the AG at the intron 3' end (Merendino, Guth et al. 1999, Wu, Romfo et al. 1999, Zorio and Blumenthal 1999). The U4/U6.U5 tri-snRNP particle is then recruited, forming the pre-catalytic spliceosome (Will and Luhrmann 2011). After the release of U1 and U4, the adenosine residue at the branch point undergoes a nucleophilic attack on the 5' ss, resulting in the formation of a 2',5'-phosphodiester bond. This is followed by the second trans-esterification reaction, which consists in the 5' ss-mediated attack on the 3' ss, giving rise to the spliced product and releasing the intron lariat (Will and Luhrmann 2011) (**Figure 3**).

The vast majority of introns are processed by the U1, U2, U4, U5, and U6 snRNP complex, also known as the major spliceosome. Yet, a minority of introns are spliced by a distinct type of snRNP complex called the minor spliceosome (Patel and Steitz 2003). Overall, the major and minor spliceosomes share many common features and the mechanism of splicing is nearly identical. However, the minor spliceosome is composed of four distinct snRNAs termed U11, U12, U4atac, and U6atac that have a counterpart in U1, U2, U4 and U6, respectively (Hall and Padgett 1996, Tarn and Steitz 1996, Tarn and Steitz 1996, Will, Schneider et al. 1999).

Most of the interactions between pre-mRNA and spliceosomal snRNAs are weak and prone to be modulated by multiple mechanisms that involve the binding of splicing regulatory proteins to the pre-mRNA, the formation of secondary structures in the pre-mRNA, the rate of RNAPII transcriptional elongation, and epigenetic modification of the template chromatin [reviewed in (Naftelberg, Schor et al. 2015)]. Depending on the combinatorial effect of factors that either enhance or repress the recognition of consensus sequences by the spliceosome, different splice sites will be selected in the pre-mRNA [reviewed in (Black 2003)]. The recent combination of large-scale characterization of alternative splicing and genome-wide identification of *in vivo* binding sites of splicing regulators unraveled the global principles guiding splicing regulation by specific RNA-binding proteins (Barash, Calarco et al. 2010, Witten and Ule 2011). Typically, splicing regulatory elements are classified as exonic or intronic splicing enhancers or silencers depending on

their location and ability to stimulate or inhibit splicing (Barash, Calarco et al. 2010, Witten and Ule 2011).

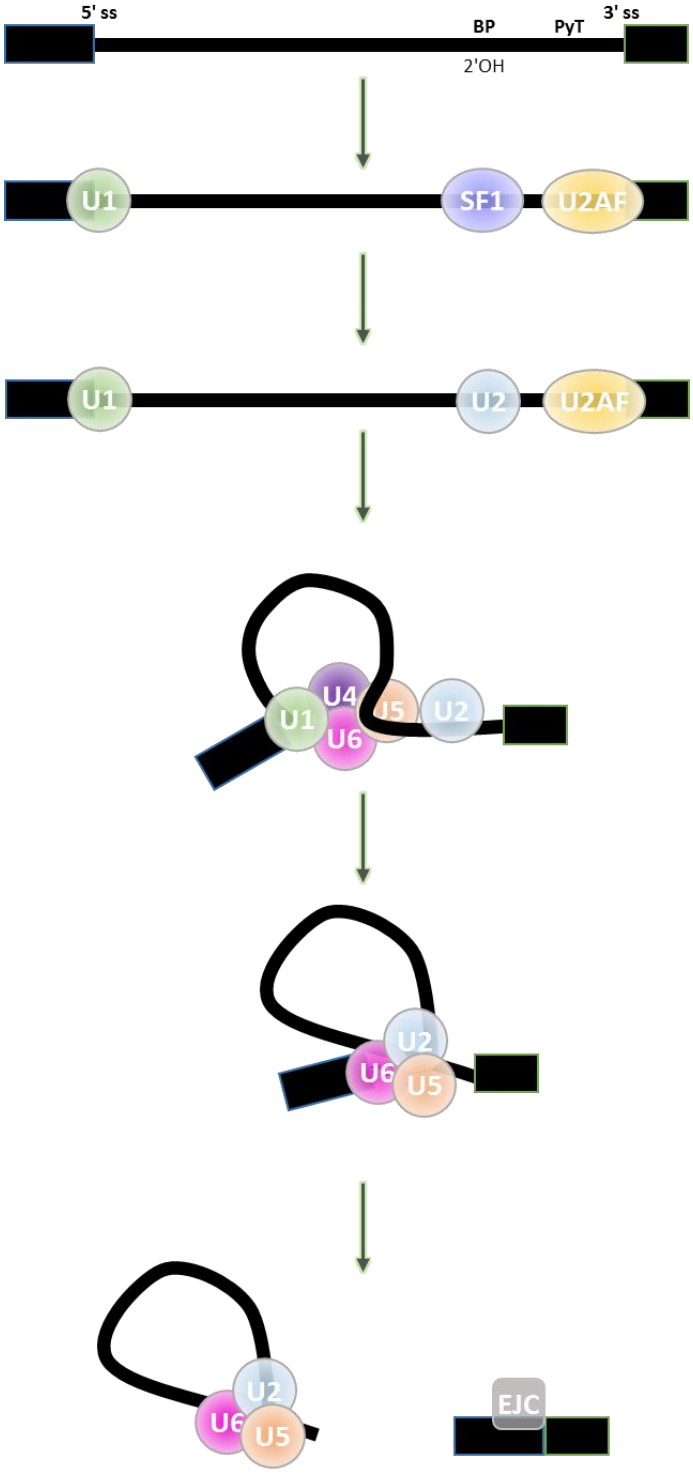


Figure 3 - *Cis*-acting RNA elements and *trans*-acting snRNPs involved in splicing.
RNA splicing is catalysed by an assembly of five snRNPs (coloured circles) and auxiliary factors, which together constitute the spliceosome. The spliceosome recognizes the splicing signals on the pre-mRNA molecule through base-pairing and catalyses the

two splicing reaction steps. After splicing, the EJC binds to the mRNA molecule. BP – branch point, ss – splice site, PyT – polypyrimidine tract, EJC – exon junction complex.

Splicing can be functionally coupled to transcription. The first evidence that introns can be removed while transcripts are still attached to the chromatin and before 3' end processing takes place was shown for a *Drosophila melanogaster* gene using "Miller spread" electron microscopy (Beyer and Osheim 1988). Electron micrograph images showed that the spliceosome formed shortly after synthesis of the 3' ss and that splicing of pre-mRNA often occurred on the nascent transcript. The idea that splicing occurs mainly co-transcriptionally is now a general consensus in the field (Brugiolo, Herzel et al. 2013, Bentley 2014). Recent genome-wide studies showed that co-transcriptional splicing frequencies are similar among different species (Brugiolo, Herzel et al. 2013). Studies performed in *S. cerevisiae* (Carrillo Oesterreich, Preibisch et al. 2010), *Drosophila* (Khodor, Rodriguez et al. 2011) and human cell lines and tissues (Ameur, Zaghlool et al. 2011, Girard, Will et al. 2012, Tilgner, Knowles et al. 2012, Windhager, Bonfert et al. 2012), using different sources of nascent transcripts (chromatin-associated and thio-labelled transcripts), showed that the frequency of co-transcriptional splicing ranges between 75-85%. Interestingly, the frequency of co-transcriptional splicing in mouse cells seems to be lower (Bhatt, Pandya-Jones et al. 2012, Khodor, Menet et al. 2012).

As in capping, the structural coupling between splicing and transcription is mediated through the RNAPII CTD domain (Custodio and Carmo-Fonseca 2016). Several evidences support the implication of the CTD in splicing kinetics. Transcription of a protein-coding gene by RNAPI or RNAPIII, that do not contain a CTD domain, impairs pre-mRNA splicing and 3' end processing (Smale and Tjian 1985, Sisodia, Sollner-Webb et al. 1987). Additionally, it was shown that a recombinant RNAPII with a truncated CTD domain was not able to recruit the splicing machinery and this was a direct cause of the observed reduction in splicing efficiency (McCracken, Fong et al. 1997, Misteli and Spector 1999). Some of the splicing factors shown to bind to the CTD domain of RNAPII are PSF (polypyrimidine tract binding protein-associated splicing factor), p54nrb/NonO (Emili, Shales et al. 2002) and U2AF65 (David, Boyne et

al. 2011). Another protein complex that is not directly implicated in splicing reaction itself but regulates the choice of the spliceosome for alternative exons is DBIRD (DBC1-ZNF326). DBIRD complex modulates transcription elongation rate and acts at the interface between nascent RNA molecules and the largest subunit of RNAPII, integrating transcript elongation with the regulation of alternative splicing (Close, East et al. 2012).

The elongation rate of RNAPII is an important property of the transcriptional machinery that can influence alternative splicing decisions by determining the duration a weak splice site endures in the pre-mRNA before a strong splice site is transcribed. More precisely, mutations in RNAPII that slow down transcription elongation rate facilitate assembly of the spliceosome at weak splice sites of alternative exons, increasing the exon inclusion/exon skipping ratio (de la Mata, Alonso et al. 2003, Nogues, Kadener et al. 2003, Luco, Allo et al. 2011, Fong, Kim et al. 2014). Additionally, RNAPII density is increased around splice sites, suggesting that RNAPII pauses for splicing to occur or that spliceosome assembly slows down RNAPII (Alexander, Innocente et al. 2010, Mayer, di Iulio et al. 2015, Mayer, Landry et al. 2017). The observation of splicing intermediates tethered to elongating RNAPII specifically phosphorylated in the Ser5 residue of the CTD, pointed that splicing occurs in association with RNAPII containing the CTD phosphorylated predominantly on Ser5 (Nojima, Gomes et al. 2015). This implies a new layer of complexity in the CTD code, in which the CTD phosphorylation pattern changes as RNAPII transcribes an intron-exon boundary. Consistent with this specific phosphorylation pattern associated with the splicing reaction, it was shown that spliceosome components are bound to the CTD when it is phosphorylated on Ser5 downstream of acceptor splice sites (Harlen, Trotta et al. 2016).

Since splicing can occur while pre-mRNA is still bound to the DNA template through RNAPII, it is reasonable to think that DNA structure may influence alternative splicing. Accordingly, it was demonstrated that binding of the transcriptional repressor CTCF to a specific intragenic unmethylated C-rich DNA sequence induces the inclusion of alternative exons in the mRNA molecules (Shukla, Kavak et al. 2011). As previously mentioned, exons have increased nucleosome occupancy compared with their flanking introns. More recently, it was proposed that histone modifications and other chromatin

features that activate transcription can be co-opted to participate in the regulation of the splicing of exons that are in physical proximity to promoter regions (Curado, Iannone et al. 2015). Other works suggested that splicing can affect chromatin organization (Keren-Shaul, Lev-Maor et al. 2013) and promote modification of histones (de Almeida, Grosso et al. 2011). In the first case, it was shown that strengthening of the 5' splice site or strengthening the base pairing of U1 snRNA with an internal exon abrogated the skipping of the internal exons and also affected chromatin organization though nucleosome remodeling (Keren-Shaul, Lev-Maor et al. 2013). In the second work, genome-wide analysis indicated that splicing is mechanistically coupled to recruitment of the methyltransferase HYPB/Setd2 to elongating RNAPII (de Almeida, Grosso et al. 2011). Overall, these studies show that there is a bi-directional interplay between the epigenetics mechanisms and RNA splicing.

Termination of transcription and 3' end processing

To produce a translatable RNA molecule and avoid read-through transcription, RNAPII and nascent transcripts must be released from the DNA at the 5' end of the template strand of a protein-coding gene, a process known as transcription termination. Release of RNA is intrinsically coupled to 3' end processing, which includes cleavage and polyadenylation of the nascent transcript. Both processes rely on conserved sequences and on the action of specific proteins (**Figure 4**).

Signals that stimulate 3' end processing (and transcription termination) are composed of a central conserved sequence motif AAUAAA (or less frequently AUUAAA) polyadenylation (pA) site, flanking auxiliary elements: a U- or GU-rich downsteam sequence element (DSE) and a U-rich upstream sequence element (USE) (Proudfoot and Brownlee 1976, Proudfoot 2011). Recognition of these regulatory sequences at the 3' end of transcripts is carried out by the cleavage and polyadenylation (CPA) complex that contains four major subcomplexes, including cleavage and polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF), cleavage factor I and II (CF Im and CF IIm) (Millevoi and Vagner 2010, Sun, Zhang et al. 2017). The site of cleavage in most pre-mRNAs lies between the pA site and the DSE, at a CA dinucleotide

and is catalyzed by CPSF subunit CPSF3 (Mandel, Kaneko et al. 2006, Shi and Manley 2015).

Pre-mRNA cleavage gives rise to two RNA molecules: an upstream pre-mRNA molecule that is stabilized by a non-templated poly(A) tail addition, the final step in mRNA biogenesis; a downstream non-coding RNA byproduct that is destabilized due to degradation by the 5'-3' exonuclease XRN2 (West, Gromak et al. 2004). Polyadenylation of the upstream molecule consists in the addition of a 50-100 nucleotides poly(A) tail (Chang, Lim et al. 2014) and it is performed by poly(A) polymerase (PAP) and a nuclear poly(A) binding protein (PABPN). 3' end processing protects RNA from degradation by 3'-5' exonucleases, stimulates export from the nucleus and translation in the cytoplasm.

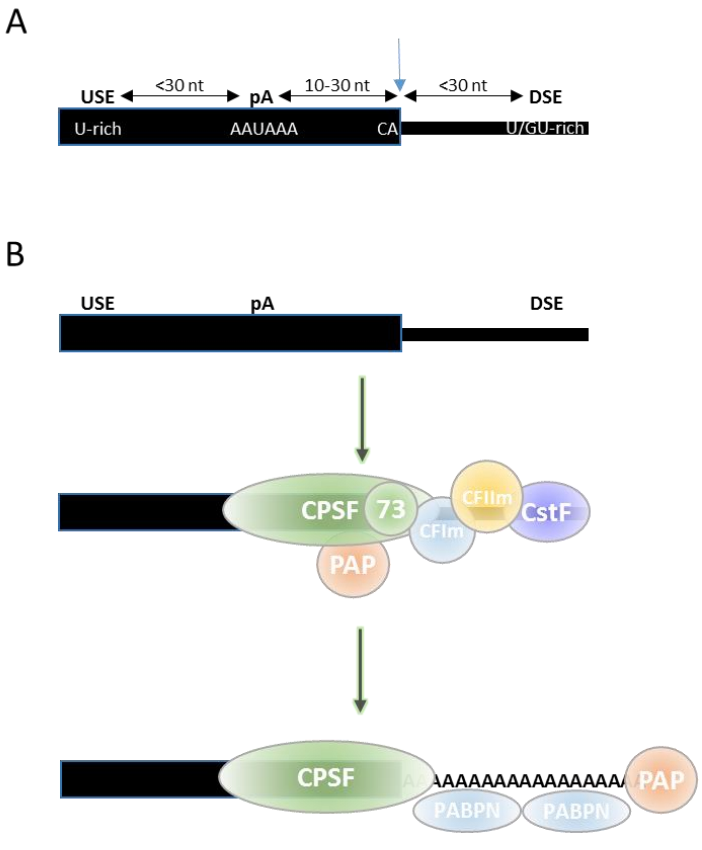


Figure 4 - *Cis*-acting RNA elements and *trans*-acting proteins involved in 3' end processing.
A) Cleavage and polyadenylation of pre-mRNA molecules requires four elements localized at 3' end of pre-mRNA molecules, an USE, DSE, pA signal and the dinucleotide CA marking the cleavage site (blue arrow). B) Those sequences are specifically recognized by four multisubunit complexes (coloured circles): CPSF, CstF,

CFI and CFII as well as PAP that is responsible for the addition of a poly(a) tail to cleaved RNAs. USE – upstream element, DSE – downstream element, pA – polyadenylation signal, CPSF – cleavage and polyadenylation specificity factor, CstF – cleavage stimulation factor, CFIm and CFIIm cleavage factor I and II, PAP – polyadenylation polymerase, PABPN – nuclear poly(A) binding protein

Moreover, 3' end processing is coupled to transcription termination, the last stage of the transcription cycle that ends with the release of RNAPII from the DNA template that is then free to restart transcription on the same or on a different promoter. This coupling relies on specific RNA sequences and modulation of the RNAPII CTD phosphorylation pattern.

Further downstream of pA signals, additional elements contribute to enhance cleavage and transcription termination processes. It has been described two categories of terminator sequences. One of such categories enhances RNAPII pausing as it transcribes the G-rich sequences element (Gromak, West et al. 2006). The other class of terminator elements may occur 1-2 kb downstream the pA site and mediate cotranscriptional cleavage (CoTC) of transcripts prior pA site cleavage. CoTC AT-rich terminator element is a common feature of around 80 human genes (Nojima, Dienstbier et al. 2013), including β-globin where it was first identified (Dye and Proudfoot 2001).

As in the other pre-mRNA processing mechanisms, the CTD phosphorylation pattern together with RNAPII pausing play a critical role in coordinating 3' end processing and transcription termination. Particularly, phosphorylation of the serine 2 residue is enriched at the end of genes and allows interaction of RNAPII with CPA factors (Heidemann, Hintermair et al. 2013). Compared with transcription initiation and elongation, kinetics of transcription termination is less understood. Currently, two models are used to explain how termination is triggered after transcription of the pA site or cleavage of the nascent RNA (Gromak, West et al. 2006, Proudfoot 2016). The first model (allosteric model), suggests that RNAPII undergoes a conformational change induced by the binding of the CPA factors that results in pausing followed by release of the transcriptional machinery (Zhang, Rigo et al. 2015). The second model (torpedo model) is based on the kinetic competition concept in which the

degradation rate of the cleaved transcript competes with the elongation rate of RNAPII. Specifically, after cleavage of the nascent transcript by the CPA complex, the nuclear 5'-3' exonuclease XRN2 is recruited to the 3' end of genes and starts degrading the cleaved downstream transcript until it reaches elongating RNAPII, which triggers its release from the DNA template (West, Gromak et al. 2004, Proudfoot 2016).

Deep-sequencing analysis has revealed that around 70% of protein-coding eukaryotic genes produce multiple mRNA isoforms with distinct 3' ends through the process of alternative polyadenylation (APA). The selective usage of alternative poly(A) sites leads to transcript isoforms with different coding potentials and/or variable 3' UTR. Shorter or longer 3' UTR sequences define which cytoplasmic pathways mRNAs are targeted to, influencing RNA stability, localization and efficiency of translation. Cellular levels of specific cleavage and polyadenylation factors, kinetics of transcription, competition between CPA factors and other RNA-binding or CPA and splicing factors are known conditions to affect APA [reviewed in (Gruber, Martin et al. 2014, Tian and Manley 2017)]. Eukaryotic cells perform reciprocal regulation of splicing and 3' end processing (Kaida 2016). The splicing factors U2AF65 and U2 were found to stimulate 3' end processing as CPSF contributes to prompt splicing of the last exon. Additionally, U1 snRNP was also found to play a role in protecting the integrity of the transcriptome by blocking the recognition of premature cleavage and polyadenylation signals that are widely spread throughout mammalian introns (Kaida, Berg et al. 2010).

Functional coupling and inter-dependence between different cellular processes allows the development of quality control mechanisms of gene expression as a way to maintain the integrity of the human transcriptome and proteome. RNA quality control/surveillance mechanisms will be further explored in the subchapter 1.4.

1.3. Functions of introns and their role in messenger RNA biogenesis

For many years the physiological importance of a genomic sequence was primarily associated with its capacity to code for proteins and therefore intronic sequences were initially assumed to be largely non-functional. However, several lines of recent evidence argue for intron functionality, as discussed in detail in the following sections.

Intron conservation

Because introns are removed from nascent transcripts during pre-mRNA processing, intronic sequences in genes have been considered as "junk DNA". However, there is a remarkable conservation of many intron positions along with highly conserved sequence elements, implying that at least some intronic features are subject to evolutionary constraints (Mattick and Gagen 2001, Hare and Palumbi 2003, Rogozin, Wolf et al. 2003). Conserved elements in introns include the consensus splice-site sequences, the binding sites for regulatory proteins, the sequences of non-coding RNA genes, as well as additional regions (Kelly, Georgomanolis et al. 2015).

Several studies revealed that first introns (i.e., introns at the 5' end of genes) are typically the longest and most conserved (Park, Hannenhalli et al. 2014). Conservation of the first intron is probably related to the presence of regulatory elements (Gaffney and Keightley 2004) and a specific pattern of chromatin organization (Bieberstein, Carrillo Oesterreich et al. 2012). Additionally, the position of introns that contain RNA genes was also found to be highly conserved (Chorev and Carmel 2013).

Introns as enhancers of transcription

Introns were first shown to increase transcriptional efficiency in transgenic mice (Brinster, Allen et al. 1988). Subsequent studies showed that intron-containing genes presented higher levels of transcription when compared to intronless genes in yeast (Juneau, Miranda et al. 2006), Drosophila (McKenzie and Brennan 1996) and mammalian cells (Brinster, Allen et al. 1988, Shabalina, Ogurtsov et al. 2010). Transcription of mammalian genes relies on a complex communication between promoters and enhancers that are often located a large distance apart in the genome, and recent studies suggest that some promoters work in combination with regulatory sequences located within introns (**Figure 5**) (Stadhouders, van den Heuvel et al. 2012). For example, expression of the type II collagen α1 (*Col2a1)* gene is dependent on SOX9, a master transcription factor that binds to regulatory regions located in *Col2a1* introns 1 and 6 (Yasuda, Oh et al. 2017). Likewise, expression of the vascular endothelial growth factor receptor *Flk1* gene requires a regulatory region located in intron 10 (Becker, Sacilotto et al. 2016). Another example is an enhancer for the Sonic Hedgehog *SHH* gene, which is located 1 Mbp upstream, within an intron of the unrelated *LMBR1* gene (Sagai, Hosoya et al. 2005). The promoter-proximal 5' splice site was further shown to stimulate transcription independently from splicing, presumably through the binding of U1 snRNP and its interaction with transcription initiation factors (Kwek, Murphy et al. 2002, Damgaard, Kahns et al. 2008).

Intron-encoded RNA genes

After splicing, introns initially excised in lariat form are first debranched (Ruskin and Green 1985) and then in most cases rapidly degraded (Sharp, Konarksa et al. 1987). Yet, not all introns are fully degraded but rather give rise to functional non-coding RNA by-products (**Figure 5**) (Mattick 2001, Hube and Francastel 2015). These include most small nucleolar RNAs (snoRNAs), which are produced from processed introns derived from genes encoding various ribosomal proteins, ribosome-associated proteins, nucleolar and other proteins (Maxwell and Fournier 1995). Remarkably, some genes have no protein-coding capacity and their primary function may be to generate

snoRNAs from their introns (Tycowski, Shu et al. 1996, Bortolin and Kiss 1998). In addition to snoRNAs, a class of unconventional micro RNAs (miRNAs) is also produced from introns. In this case, pre-miRNA-like hairpins are generated by the spliceosome followed by lariat-debranching and exosome mediated trimming (Flynt, Greimann et al. 2010). These atypical miRNA precursors are called mirtrons due to their location in introns from protein coding and non-coding genes (Berezikov, Chung et al. 2007, Okamura, Hagen et al. 2007, Valen, Preker et al. 2011).

Alternative splicing and intron retention

Alternative splicing increases transcriptome and proteome diversity by generating multiple mRNA isoforms from a single gene. A pre-mRNA molecule can be alternatively spliced through exon skipping, alternative splice site selection, and intron retention [reviewed in (Black 2003, Chen and Manley 2009)]. For many years, intron retention in mature mRNAs was considered a consequence of mis-splicing, as intron-containing mRNAs are often targeted for degradation by the exosome in the nucleus or nonsense-mediated decay in the cytoplasm (Jaillon, Bouhouche et al. 2008, Roy and Irimia 2008, Gudipati, Xu et al. 2012). However, recent transcriptomic analysis revealed that many introns are actively retained in polyadenylated transcripts and contribute to downregulate gene expression (Yap, Lim et al. 2012, Wong, Ritchie et al. 2013). Yet, transcripts with retained introns are not necessarily short-lived. For example, in the mouse brain, mRNAs containing certain introns are stably accumulated in the nucleus, but in response to a stimulus, these molecules are spliced and acutely transported to the cytoplasm (Mauger, Lemoine et al. 2016, Naro, Jolly et al. 2017). Similarly, nuclear accumulation of stable transcripts containing retained introns was detected at specific stages during spermatogenesis (Naro, Jolly et al. 2017). The term “detained” introns has been proposed to describe this class of introns that are transiently retained in nuclear transcripts but can still be spliced (Boutz, Bhutkar et al. 2015). Retained introns can additionally affect gene expression by other mechanisms. Namely, the first intron of the ZEB2 pre-mRNA contains an internal ribosome entry site, so that retention of this intron allows more efficient translation (Beltran, Puig et al. 2008), and retention of the third intron in the rat Ceacam6-

L pre-mRNA generates a novel protein isoform in male germ cells (Kurio, Murayama et al. 2008).

Non-canonical splicing

Recent developments in transcriptome sequencing and analysis have revealed a remarkable prevalence of unconventional or non-canonical splicing mechanisms ranging from recognition of atypical splice sites to changes in the usual order of splicing [reviewed in (Sibley, Blazquez et al. 2016)].

Many sequences similar to the consensus motifs of canonical splice sites are present throughout introns. These sequences are known as cryptic, non-canonical or pseudo splice sites (**Figure 5**). What determines preference for a bona fide versus a pseudo splice site is still unclear, particularly after the finding that actual splice site sequences can be extremely diverse (Roca, Sachidanandam et al. 2003, Roca, Krainer et al. 2013). Indeed, over 9000 sequence variants have been found in the -3 to +6 region of human 5' splice sites (Roca, Akerman et al. 2012), challenging the dogma that spliceosome recognition relies primarily on consensus sequences at exon-intron boundaries. Moreover, cryptic splice sites are often used when a natural splice site is mutated, further arguing that splice site recognition is not intrinsically defined by any given sequence (Roca, Sachidanandam et al. 2003, Roca, Krainer et al. 2013). Most likely, bona fide splice site selection results from the combinatorial effect of proteins such as SR and hnRNP proteins that bind to the pre-mRNA and either stabilize spliceosome interactions or inhibit the recruitment of spliceosomal components (Liu, Zhang et al. 1998, Dreyfuss, Kim et al. 2002).

Intronic sequences that are flanked by non-canonical splice sites and are normally not observed in spliced mRNAs are referred to as pseudo-exons or cryptic exons. Compared to genuine exons, pseudo-exons tend to have less splicing enhancer and more splicing silencer motifs (Sironi, Menozzi et al. 2004, Wang, Rolish et al. 2004, Zhang and Chasin 2004, Corvelo and Eyras 2008). Pseudo-exons often derive from transposable elements, in particular from antisense *Alu* sequences [reviewed in (Keren, Lev-Maor et al. 2010)].

A non-canonical mechanism of intron removal referred to as recursive splicing was first detected in long Drosophila pre-mRNAs (Hatton, Subramaniam et al. 1998, Burnette, Miyamoto-Sato et al. 2005). Recently, recursive splicing was also observed in long introns of mammalian brain-specific transcripts (Sibley, Emmett et al. 2015). These introns contain a cryptic site termed a recursive splice site or a "zero-length exon" consisting in a combination of 3′ and 5' splice sites that allow an intron to be spliced in multiple consecutive steps (Duff, Olson et al. 2015, Sibley, Emmett et al. 2015). In this process, the 3' splice site is used to splice the upstream part of the intron, which reconstitutes a 5' splice site that is then used to splice the downstream part. Evidence for multi-step recursive splicing of dystrophin pre-mRNAs in human skeletal muscle cells has also been reported (Gazzoli, Pulyakhina et al. 2016).

In some cases, the pre-mRNA splicing reaction does not follow its canonical order but rather occurs in a reversed orientation that links a downstream 5' (donor) site to an upstream 3' (acceptor) site to produce a circular RNA (for recent reviews see (Chen 2016, Salzman 2016). To date, thousands of circular noncoding RNAs generated by "backsplicing" of transcripts from protein-coding genes have been reported. Circularization, which results, for example, from covalently joining the two ends of a single exon, can be favoured by the presence of inverted repeats, such as *Alu* elements, in the flanking introns (Jeck, Sorrentino et al. 2013, Liang and Wilusz 2014, Wilusz 2015, Dong, Ma et al. 2016). Intronic circular RNAs have further been detected resulting from intron lariats that are resistant to de-branching due to C-rich motifs surrounding the branch point (Zhang, Zhang et al. 2013). Although many circular RNA species appear to result from splicing errors, some may function as modulators of gene expression [reviewed in (Chen 2016, Salzman 2016)].

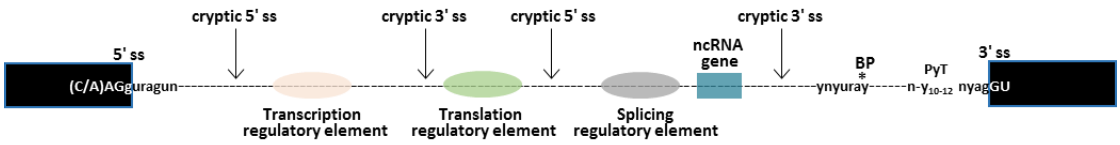


Figure 5 – Functional elements located deep within introns.
Introns contain different classes of functional elements such as cryptic splice sites, ncRNA genes and binding sites for transcription, splicing and translation regulatory elements.

1.4. Quality control of messenger RNA biogenesis

The expression level of a certain RNA is determined by the rate of synthesis and rate of degradation. In humans, there are three RNAPs responsible for polymerizing RNA, while degradation of such molecules is carried out by more than sixty nucleases (Stoecklin and Muhlemann 2013).

Nucleases are important determinants of the steady state levels of RNAs in a cell and constitute an important way of gene expression regulation. One of such group of enzymes is enrolled in processing steps described in the previous subchapter: splicing; 3' end processing and transcription termination. Additionally, ribonucleases can be involved in gene expression regulation, host defense and RNA quality control (QC) [reviewed in (Ghosh and Jacobson 2010)]. They can operate as endoribonucleases, which cleave RNA polymers internally, and as exoribonucleases, which remove nucleotides one at a time from either the 3' or the 5' end of the RNA molecule.

Quality control mechanisms play an important cellular role avoiding the accumulation of misprocessed, nonfunctional mRNAs. Acting in both the nucleus and the cytoplasm, multiple quality control pathways monitor degradation, arrest and/or downregulation of production of potentially harmful molecules (Houseley and Tollervey 2009, Ghosh and Jacobson 2010, Schmid and Jensen 2010, Muhlemann and Jensen 2012, Porrua and Libri 2013).

Nuclear quality control in human cells

Nuclear QC mechanisms have been mainly described in yeast, however there are important works describing RNA surveillance mechanisms in human cells.

The multisubunit RNA exosome complex is the major ribonuclease of eukaryotic cells that participates in the quality control of mRNAs [reviewed in (Kilchert, Wittmann et al. 2016)]. The human exosome is composed of nine

subunits arranged in a two-layered ring wide enough to accommodate single-stranded RNA molecules. Its ribonucleolytic subnunits RRP44 and RRP6 confer endonuclease (RRP44) and 3'-5' exonuclease (RRP44 and RRP6) activities that are essential to degrade defective RNA molecules, releasing nucleoside 5'-monophosphates. All components of the human exosome are distributed throughout the nucleolus, nucleoplasm and cytoplasm (Kilchert, Wittmann et al. 2016).

Correctly processed mRNAs display secondary structures and exist in the cell in association with different proteins in ribonucleoprotein (RNP) complexes. These features act as protection to the action of the nuclear exosome. Errors during pre-mRNA processing, often lead to alterations in the secondary structures and in the formation of RNPs, which lead defective RNAs more susceptible to be degraded by the exosome [reviewed in (Belair, Sim et al. 2017, Bjork and Wieslander 2017)].

The exosome acts as an effector, as it is rarely involved in the decision process that commits a molecule for degradation. Instead, it relies on auxiliary factors that direct the exosome to its substrates, guaranteeing target specificity and the consequent removal of the intended molecules (Zinder and Lima 2017).

In the nucleus, the helicase co-factor hMTR4/SKIV2L2 facilitates substrate recognition through RNA-binding protein adaptors. Human MTR4 has been identified as part of three nuclear complexes: human Trf4p/5p-Air1p/2p-Mtr4p polyadenylation (hTRAMP), nuclear exosome targeting (NEXT) and poly(A) tail exosome targeting (PAXT) (Zinder and Lima 2017) (**Figure 6**).

Human TRAMP complex is localized in the nucleolus and is involved in oligoadenylation of nucleolar rRNAs to facilitate their decay. RNA length and polyadenylation status are biochemical parameters discriminating whether a nucleoplasmic exosome substrate is likely to be targeted by the NEXT or the PAXT pathways. NEXT is composed of hMTR4, the Zn-finger protein ZCCHC8, and the RNA-binding factor RBM7 and primarily targets early and unprocessed transcripts (Lubas, Andersen et al. 2015). An iCLIP genome-wide analysis showed that RBM7 targets preferentially pre-mRNA relative to mRNA and accumulates at intron 3' ends with a clear enrichment for the cap-proximal 1000 nucleotides (Lubas, Andersen et al. 2015). Moreover, binding of RBM7

does not always result in pre-mRNA degradation by the human exosome, suggesting a kinetic coupling between processing and degradation.

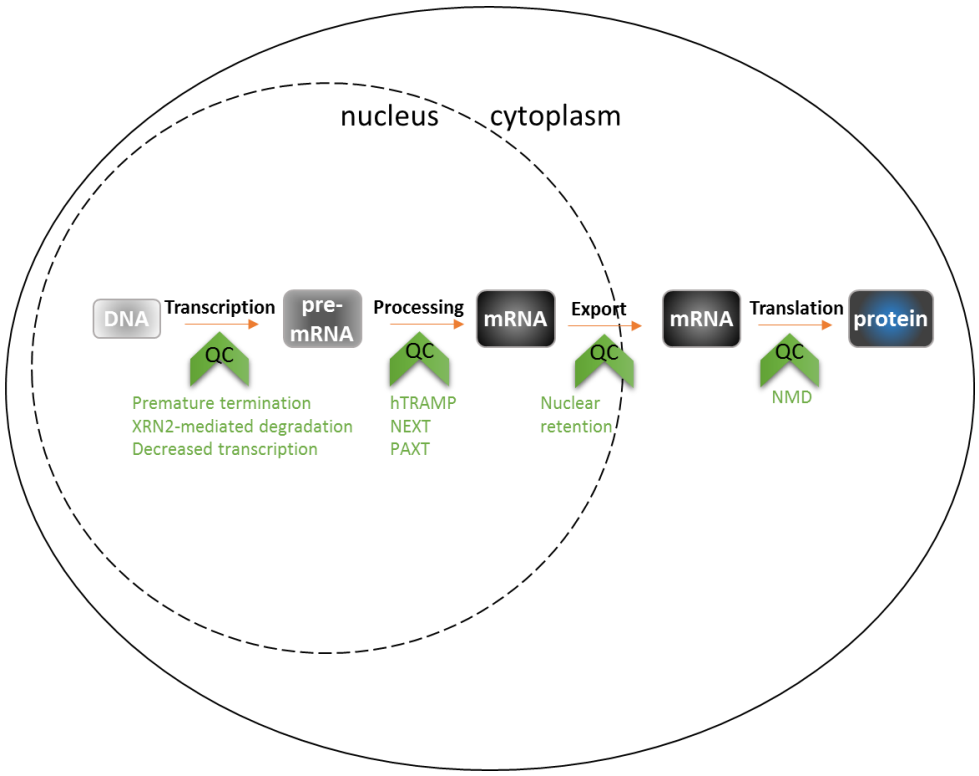


Figure 6 – RNA quality control mechanisms in mammalian cells.
Aberrant RNA molecules can be destabilized both in the nucleus and cytoplasm of human cells. Different quality control (QC) mechanisms operate to prevent the accumulation of defective RNAs either by targeting them for co- or post-transcriptional nuclear degradation, blocking their export to the cytoplasm, or avoiding their translation through NMD cytoplasmic degradation. QC – quality control, NMD – nonsense mediated decay.

PAXT comprises the ZFC3H1 Zn-knuckle protein as a central link between hMTR4 and the nuclear poly(A)-binding protein PABPN1. PAXT promotes degradation of longer and more extensively polyadenylated substrates compared with NEXT substrates (Meola, Domanski et al. 2016).

Different evidences support the idea that the exosome can play a role in co-transcriptional quality control: both NEXT and PAXT interact with the cap-binding complex containing ARS2 via an adaptor protein, ZC3H18, physically tethering the exosome to nascent capped transcripts to promote degradation following termination (Winczura, Schmid et al. 2018); nuclear exosome can

interact with human spliceosome components, suggesting a physically link between RNA decay and pre-mRNA splicing to ensure generation of properly spliced RNA and decay of aberrant transcripts (Nag and Steitz 2012).

Processing defects can indeed trigger defective transcripts for degradation by the exosome. One such example is the arrest of splicing defective β-globin transcripts in the vicinity of the gene locus in an exosome-dependent manner (Custodio, Carmo-Fonseca et al. 1999, de Almeida, Garcia-Sacristan et al. 2010). Retention of splicing defective transcripts at the transcription site showed that errors in splicing render RNAPII incompetent for transcription termination, while normally processed transcripts are rapidly released to nucleoplasm and exported to the cytoplasm (Martins, Rino et al. 2011).

Moreover, an intronless β-globin reporter that is inefficiently exported to the cytoplasm, as well as viral nuclear noncoding RNA and some endogenous lncRNAs are targeted for degradation by the exosome in a PABPN1-dependent manner (Bresson and Conrad 2013). Interestingly, this pathway targets intronless β-globin but does not degrade spliced β-globin reporter. In addition to PABPN1, targeting of intronless RNAs to PAXT requires poly(A) polymerase-dependent extension of the poly(A) tail (Meola, Domanski et al. 2016), suggesting that extension of the poly(A) tail provides the exosome a binding site, which is in accordance with the observation that binding of the exosome relies on the accessibility of around 30 nucleotides of “naked” RNA (Bresson and Conrad 2013).

Termination of transcription can be used as a QC mechanism as a way to avoid the production of defective RNAs and this process is tightly connected to exosome activity. Premature transcription termination can be induced in response to transcription errors, such as DNA damage or mis-synthesis of RNA molecules. It is essential in controlling pervasive transcription as well as supressing bidirectionality of promoters (Proudfoot 2016). Pervasive transcription gives rise to a class of non-coding, short and divergently originated from canonical promoters transcripts. These transcripts, named PROMPTs, present low steady state levels since they are rapidly triggered to degradation by the human exosome (Core, Waterfall et al. 2008). Directionality of promoters is driven by the differential presence of pAs and 5' ss. Regions upstream of human bidirectional promoters are markedly enriched in pAs,

while regions downstream of promoters are enriched in 5' ss. Promoter-proximal 5' ss were shown to prevent the usage of pAs, avoiding premature termination of transcription (Kaida, Berg et al. 2010, Almada, Wu et al. 2013). The mechanistic reason that PROMPTs are unstable is related to the pAs that are present more frequently in the non-functional transcripts. Those 3' end signals are recognized by CPSF and CF factors inducing promoter proximal termination. Additionally, it was shown that the interaction of CBC with ARS2 can contribute to PROMPTs termination by recruiting cleavage factors and the NEXT complex (Iasillo, Schmid et al. 2017).

In addition to the nuclear exosome, the DXO and XRN families of 5'-3' exoribonucleases are critical for ensuring the fidelity of cellular RNA turnover in eukaryotes. With its decapping and 5'-to-3' exoribonuclease activities, DXO plays a central role in a pre-mRNA 5' end capping quality control mechanism that targets pre-mRNAs that fail to acquire the correct cap structure. Incompletely capped pre-mRNAs are inefficiently spliced and cleaved, suggesting a link between proper 5' end capping and subsequent pre-mRNA processing (Jiao, Chang et al. 2013).

Highly conserved across species, the XRN family is typically represented by one cytoplasmic enzyme (XRN1) and one or more nuclear enzymes (XRN2 and XRN3) (Nagarajan, Jones et al. 2013). In the nucleus, XRN2 recognizes single-stranded RNA with a 5' terminal monophosphate and degrades it to mononucleotides. It plays a central role in RNA decay, gene silencing, rRNA and snoRNA maturation, transcription termination, R-loop resolution, DNA damage signalling and repair (Miki and Grosshans 2013, Eberle and Visa 2014, Fong, Brannan et al. 2015, Kilchert, Wittmann et al. 2016, Morales, Richard et al. 2016). Additionally, it was suggested that decapping of PROMPTs and torpedo termination are also mechanisms involved in transcription termination and destabilization of PROMPTs (Lubas, Andersen et al. 2015). XRN2 has also been involved in the co-transcriptional degradation of reporter β-globin transcripts with splicing and 3' end processing defects (Davidson, Kerr et al. 2012). Similarly, some nascent endogenous pre-mRNA transcripts were shown to be stabilised upon XRN2 depletion following use of the splicing inhibitor, Spliceostatin A (SSA) (Davidson, Kerr et al. 2012).

Despite that, defective RNA molecules may be released from the chromatin and escape co-transcriptional degradation by the nuclear exosome or XRN2. Indeed, several studies reported that global splicing inhibition, either by depleting spliceosome components or upon SSA treatment, causes accumulation of processing defective mRNAs in the nucleus within nuclear speckles (Wegener and Muller-McNicoll 2017). One possible mechanism that may promote nuclear retention of unspliced transcripts is the presence of a 5' ss motif in the mature RNA molecules (Lee, Akef et al. 2015).

Coupling between alternative splicing and nuclear export has been increasingly recognized as an import mechanism of regulation of gene expression in the mammalian nervous system (Yap and Makeyev 2013). One of such examples is the alternative exon definition of the minor spliceosome component *U11/U12-65K*. Alternative splicing of the *65K* pre-mRNA generates a productive short-3' UTR splicing isoform or a non-productive long-3' UTR splicing isoform. For the short isoform, optimal terminal exon definition ensures efficient nuclear export, while for the long isoform, aberrant 3' end processing together with binding of U11/U12 snRNP to a ultraconserved sequence promote retention of the transcript (Verbeeren, Verma et al. 2017).

Interestingly, it was recently shown that a small fraction of the pre-mRNA molecules can escape nuclear retention and degradation and be exported to the cytoplasm following drug treatment with SSA, meamycin or pladienolide B (Carvalho, Martins et al. 2017). These compounds target the SF3b subunit of the spliceosomal U2 snRNP and have been used as anti-cancer drugs. Once in the cytoplasm, those transcripts are degraded by the most characterized RNA quality control mechanism in eukaryotic cells, nonsense mediated decay (NMD).

Cytoplasmic quality control in human cells

NMD is a cytoplasmic mRNA quality control that operates in all eukaryotes and is responsible for degrading transcripts harboring premature termination codons (PTC) (Kurosaki and Maquat 2016) (**Figure 6**). PTC-containing transcripts may arise from heritable nonsense and frameshift mutations in the

germline or they can be generated by routine errors in transcription and splicing. Frameshift mutations are an important source of PTCs, by disrupting or creating non-canonical splice sites or affecting exonic/intronic splicing regulators, since retention of introns in the mature mRNA can either introduce an intron-encoded PTC or induce a frameshift leading to the formation of an exon-derived PTC. Transcription error rate by RNAPII is assumed to be approximately 10^{-5}. Based on that it was proposed that 0.05%-0.5% of transcripts of any given gene are expected to contain PTCs due to mis-transcription (Cusack, Arndt et al. 2011). NMD mitigates the negative effects of PTC-containing transcripts originated from transient transcriptional errors.

Furthermore, NMD alters the expression of around 10% of different types of cellular mRNAs in response to environmental changes, the so-called endogenous NMD substrates (Ottens and Gehring 2016, Nickless, Bailis et al. 2017). Endogenous transcripts can have NMD-eliciting features at various positions, including upstream open reading frames in the 5' UTR, introns in the 3' UTR, long 3' UTR and selenocysteine codons (Mendell, Sharifi et al. 2004).

NMD must accurately distinguish a PTC from a normal stop codon on an mRNA and recruit enzymes to degrade the defective transcript. After pre-mRNA splicing an exon-junction complex (EJC) is deposited 20-24 nucleotides upstream of an exon-exon junction. EJCs upstream of and within mRNA coding regions are removed by ribosomes during the first round of translation in the cytoplasm (Le Hir, Sauliere et al. 2016). However, because PTCs shorten the length of the coding region, any downstream EJCs that normally reside within the coding region would fail to be removed from what becomes the 3' UTR (Ottens and Gehring 2016). An alternative mechanism that triggers RNAs without splicing downstream from the PTC to NMD was described. This 3' UTR EJC-independent NMD reflects the importance of the distance between a termination codon and the mRNA poly(A) tail, since it is activated by a long mRNA 3' UTR (Metze, Herzog et al. 2013). However, predicting whether an mRNA is an NMD target cannot be made based on its 3' UTR length alone, since it contains stabilizing elements (Toma, Rebbapragada et al. 2015).

Degradation of NMD substrates is initiated by the central NMD factor, the RNA helicase UPF1 (Kurosaki and Maquat 2016). Translation modulates binding of human UPF1 to cellular RNAs, however UPF1 was shown to promiscuously bind

to RNAs independently of translation. Subsequent phosphorylation of UPF1 induces conformational changes that increase UPF1 affinity for mRNA. Depending on the transcript-specific architecture, phosphorylated UPF1 can either recruit the endonuclease SMG6 or the deadenylation-promoting SMG5/7 complex: NMD substrates with PTCs undergo constitutive SMG6-dependent endocleavage, while turnover of NMD substrates containing uORFs and long 3' UTRs involves both SMG6- and SMG7-dependent endo- and exonucleolytic decay, respectively (Ottens, Boehm et al. 2017). The resulting endocleaved, decaped or deadenylated pieces or mRNA are subsequently degraded by the exosome or 5'-3' exonuclease XRN1(Muhlemann and Lykke-Andersen 2010).

Although NMD is a highly efficient post-transcriptional quality control mechanism that detects and destroys aberrant mRNAs containing PTCs [reviewed in (Popp and Maquat 2013)], additional cytoplasmic mRNA decay pathways ensure cell homeostasis. In particular, the presence of 3' end elements has been associated to regulate the levels of particular RNAs that, if accumulated, may induce toxicity for the cell. Mammalian RNAs that resulted from aberrant processing and contain stable secondary structures are targeted for TUTase-DIS3L2 surveillance (TDS) pathway. This cytoplasmic quality control involves uridylation of various cellular RNA species at the 3' end by TUT4/7 as a mark for degradation by the exosome-independent 3'-5' DIS3L2 exoribonuclease (Ustianenko, Pasulka et al. 2016). Interestingly, nuclear surveillance involves oligoadenylation tagging as a mark for degradation, whereas cytoplasmic utilizes oligouridylation. Prematurely terminated mRNA transcripts that escape nuclear degradation, as well as non-coding transcripts, such as rRNA, snRNA, tRNA, lncRNA are targeted by TDS in the cytoplasm. Uridylation has also been implicated in degradation of histone mRNAs upon completion of DNA replication.

Moreover, there are various pathways of cytoplasmic mRNA decay that are conditionally used to regulate gene expression, namely Staufen 1 (STAU1)-mediated mRNA decay (SMD), ARE- and GRE- mediated decay and micro (mi)RNA-mediated mRNA decay, Staufen (STAU)-mediated mRNA decay targets particular newly synthesized mRNAs that contain a STAU binding site (SBS) within their 3' UTR. Both STAU1 and STAU2 interact directly with the NMD factor UPF1, enhancing its helicase activity to promote effective SMD (Park and

Maquat 2013). Recently it was found that primate-specific *Alu* short interspersed elements (SINEs) can promote (SMD) when present in mRNA 3' UTRs (Lucas, Lavi et al. 2018).

A common *cis* element that confers instability to an mRNA is the AU-rich element (ARE). It was identified in the 3' UTR of 5–8% of human mRNAs, encoding for proteins of diverse functions involved in immune response and inflammation, cell cycle and carcinogenesis (Labno, Tomecki et al. 2016). Several ARE-binding proteins (ABPs) regulate mRNA stability and translation through their interaction with AREs. Less frequent, is the family of GU-rich elements (GREs), found in the 3' UTR of many human mRNAs that are involved in processes like cell growth, migration and apoptosis, conferring transcript instability (Borbolis and Syntichaki 2015, Labno, Tomecki et al. 2016). Another type of cytoplasmic decay mechanism involves the recognition of specific *cis* elements on the target mRNA by non-coding miRNAs and regulates the steady-state transcript levels of a large number of genes by promoting degradation.

1.5. Defects in messenger RNA biogenesis and human disease

Inappropriate biogenesis of mRNAs can have a tremendous impact on human health. Shortly after the discovery of splicing it was found that patients with β-thalassemia failed to produce β-globin (HBB) protein due to a point mutation in a splice site that disrupted the normal processing of *HBB* pre-mRNA (Busslinger, Moschonas et al. 1981, Spritz, Jagadeeswaran et al. 1981). Since then, alterations in pre-mRNA splicing were increasingly recognized as responsible for monogenic disorders (Krawczak, Thomas et al. 2007, Padgett 2012, Singh and Cooper 2012, Pedrotti and Cooper 2014, Sterne-Weiler and Sanford 2014, Chabot and Shkreta 2016, Scotti and Swanson 2016), as well as for complex human traits (Heinzen, Ge et al. 2008, Yu, Maroney et al. 2008). Currently estimates indicate that approximately 30% of all disease-causing mutations are assumed to disrupt splicing. Most are placed in exon-intron boundaries, splicing regulatory motifs and in core and auxiliary constituents of the spliceosome. However, the recent introduction of whole-genome sequencing approaches in clinically oriented screening studies has resulted in the identification of an increasing number of pathogenic variants located deep within introns (i.e., more than 100 base pairs away from exon-intron boundaries) that affects pre-mRNA splicing profiles [reviewed in (Scotti and Swanson 2016, Vaz-Drago, Custodio et al. 2017)].

Disease-causing mutations localized in exon-intron boundaries

Approximately 15% of disease-causing mutations in the Human Gene Mutation Database (Stenson, Mort et al. 2014) alter the consensus splice site sequences. The phenotypic outcome of these mutations can be summarized in three distinct categories: 1) Constitutive exon skipping or intron retention; 2) altered inclusion/exclusion ratio of alternative exons; and 3) activation of cryptic splice sites, resulting in inclusion/exclusion of sequences in a spliced mRNA (**Figure 7A, B, C**). Most often the final result is loss of gene function due to

either synthesis of a nonfunctional protein or disruption of the reading frame that introduces a PTC and targets the mRNA for degradation by NMD (Singh and Cooper 2012). By degrading these aberrant transcripts, NMD acts in preventing the production of truncated proteins and accumulation of misfolded protein that otherwise would have a potential dominant-negative effect on the cell [reviewed in (Miller and Pearce 2014)]. Besides splicing mutations, PTC-containing transcripts likely appear due to nonsense and frameshift mutations.

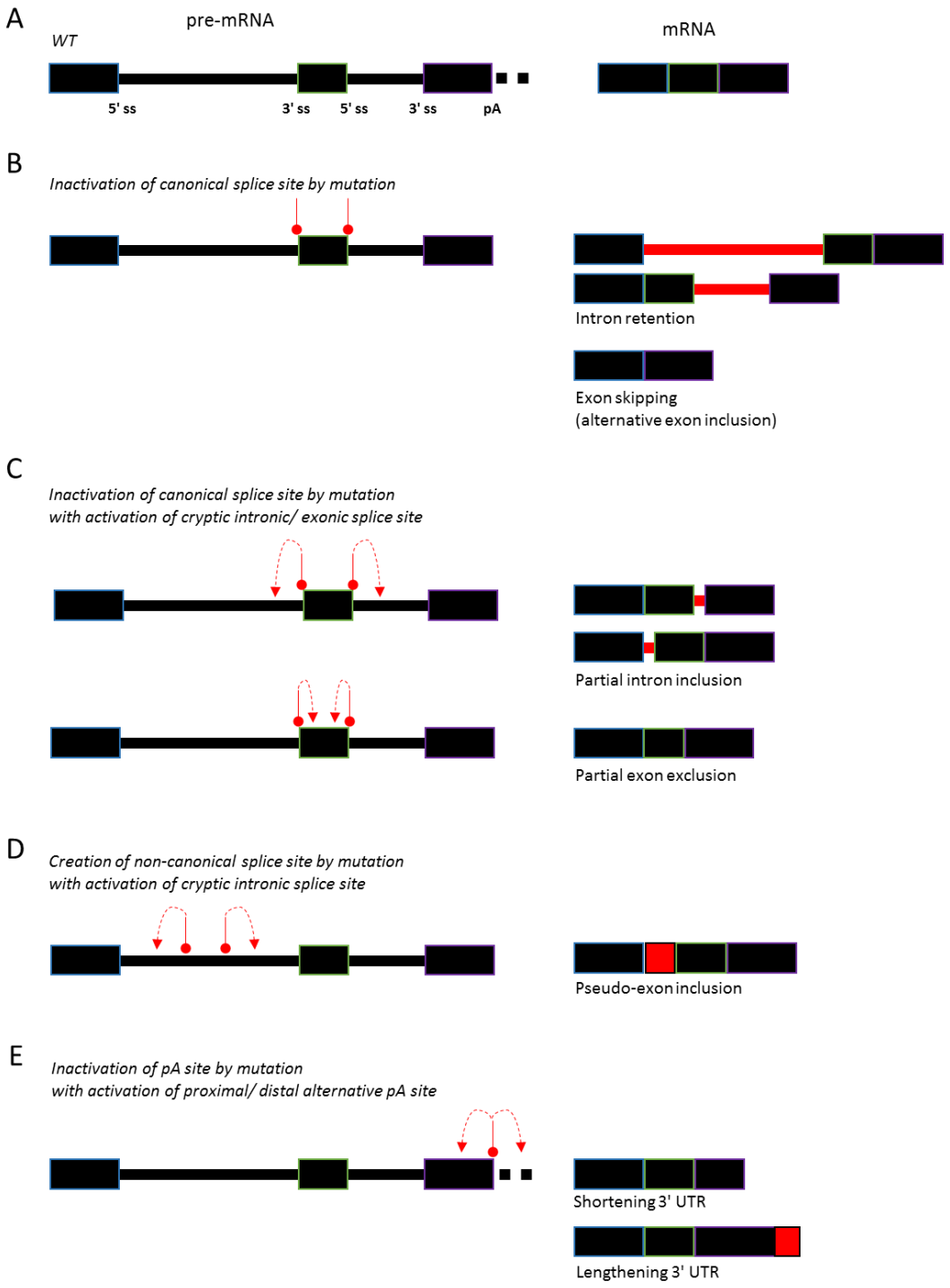


Figure 7 - Mis-processing of a single gene can result in human disease.
Different effects of aberrant splicing and polyadenylation on the usage of pre-mRNA sequences due to *cis*-acting mutations are represented. A) Healthy splicing pattern. B) Inactivation of a canonical splice site by mutation may lead to intron retention or exon skipping. C) If a cryptic splice site is activated as consequence of a mutation in a canonical splice site, intronic sequences may be included or exonic sequences excluded. D) Deep-intronic mutations usually create a non-canonical splice site with subsequent activation of a pre-existing splice site which leads to the inclusion of a pseudoexon in the final mRNA molecule. E) Mutations that inactivate pA signals can also lead to the activation of alternative pA sites and transcription termination defects. Red lines denote possible positioning of *cis*-acting mutations and red dotted lines depicted activation of cryptic pre-mRNA signals. Left panels illustrate pre-mRNA molecules while right panel illustrate the effects of healthy/aberrant splicing on mRNA molecules.

NMD is intimately linked to human health and disease, since it modulates the manifestation and clinical severity of one-third of genetic disorders (Miller and Pearce 2014). NMD can modulate the manifestation of genetic disease by altering the pattern of inheritance for a specific allele. Differential recognition and degradation of mutated transcripts can occur due to the location of the PTC (Nagy and Maquat 1998). PTC-containing transcripts are normally recognized based on the position of the translation termination codon relative to the last exon-exon junction [reviewed in (Kervestin and Jacobson 2012)]. If the PTC is located more than 50 nucleotides upstream of an exon-exon junction region, NMD will degrade the transcript, avoiding the production of a potentially toxic truncated peptide. This means that a heterozygous carrier of the mutated gene can still rely on the wild type allele for proper function, leading to an autosomal recessive pattern of inheritance, this is the case of β-thalassemia and Retinitis pigmentosa (**Figure 8**). If the PTC is located less than 50 nucleotides upstream of the last exon-exon junction or within the last exon, NMD is not triggered, allowing the truncated peptide product to accumulate in the cell, which can lead to an autosomal dominant pattern of inheritance, this can happen in some cases of β-thalassemia and spinal muscular atrophy (**Figure 8**). This means that the position-dependent recognition of the PTC within the gene can cause distinct clinical severity of a certain genetic disease or distinct traits of manifestation due to the variable involvement of NMD (Kurosaki and Maquat 2016).

Although NMD is a highly efficient cytoplasmic quality control mechanism that detects and destroys aberrant mRNAs containing PTCs (Popp and Maquat 2013), additional nuclear surveillance pathways ensure transcriptome fidelity in the context of human disease. Disease-causing mutations in splice sites can trigger retention of transcripts in the cell nucleus, preventing their export to the cytoplasm (Wegener and Muller-McNicoll 2017).

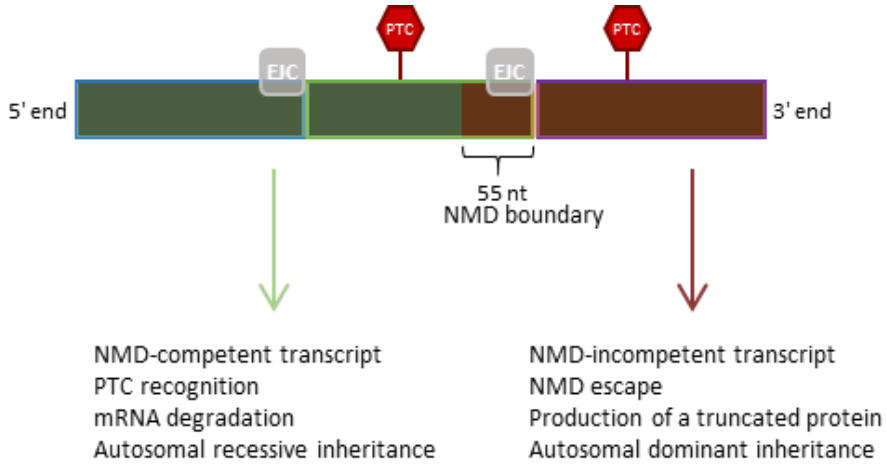


Figure 8 - Position-dependent effects of nonsense mutations of NMD correlate with inheritance pattern and clinical severity.
Differential recognition and degradation of mutated transcripts can occur due to the location of the PTC either upstream of the NMD boundary (NMD-competent) or downstream of the NMD boundary (NMD-incompetent). In mammals, the EJC serves as the downstream marker to determine the competence of a certain PTC-containing transcript to be targeted by NMD. Generally, disease-causing mutations that introduce a PTC 50-55 nt upstream of the 3' terminal exon junction elicit strong NMD responses and are the cause of autosomal recessive condition. On the other hand, disease-causing mutations that introduce a PTC downstream of the NMD boundary does not elicit NMD response and are the cause of a variety of autosomal dominant conditions. Adapted from (Holbrook, Neu-Yilik et al. 2004, Miller and Pearce 2014).

Osteogenesis imperfecta (OI) is a group of genetic disorders caused by mutations in the *COL1A1* gene, including splice-site mutations. One of such mutations affects splicing of intron 26, leading to intron retention. Analysis of OI type 1 patient-derived cells revealed that transcripts with retained intron 26 are prevented from nuclear export, which leads to decreased collagen expression (Johnson, Primorac et al. 2000).

Regulated intron retention also modulates mRNA export from the nucleus. One such example is provided by apolipoprotein E (ApoE), a stress response protein that has been described as the main susceptibility gene for Alzheimer's and other neuro-degenerative diseases. In degenerating neurons, ApoE mRNA retains intron 3 under normal conditions, which arrests the mRNA export from the nucleus to the cytoplasm and limits the ApoE protein production output (Xu, Walker et al. 2008). However, induced neuronal injury (excitotoxic stress) stimulates splicing of the partially unspliced transcript, thus allowing accumulation of translation-competent *APOE* mRNA in the cytoplasm. In contrast, in response to excitotoxic stress, *APOE-I3* expression increased in neurons with normal morphology (Xu, Walker et al. 2008).

Similar to *APOE* mRNA, intron 3 of *CYR61* transcript is retained in normal cells and skipped in various cancer types leading to high expression of CYR61, which is tightly linked to tumour progression (Hirschfeld, zur Hausen et al. 2009).

Disease-causing mutations localized deep within introns

Genome Wide Association Studies have identified many single nucleotide variants located deep within introns with significant association to diseases (Xiong, Alipanahi et al. 2015, Hsiao, Bahn et al. 2016). These findings are fostering a new era of research focused on understanding how variation in deep intronic sequence affects pre-mRNA splicing and contributes to disease phenotypes. Current estimates indicate that 1-5% of disease-causing splicing mutations are placed deep within introns (Vaz-Drago, Custodio et al. 2017). The phenotypic outcome of these mutations varies according to the sequence and function of the affected intronic region: 1) inclusion of a deep-intronic (pseudo-exon) or intronic piece in the mRNA; 2) disruption of a non-coding RNA; 3) deregulation of transcription of the mutant gene (Vaz-Drago, Custodio et al. 2017).

Pseudo-exon inclusion is now considered a more frequent cause of disease than previously thought (Dhir and Buratti 2010, Romano, Buratti et al. 2013). This aberrant process can be triggered by intronic mutations that activate non-

canonical splice sites (**Figure 7A, D**). The appearance of a pseudo-exon generally disrupts the reading frame introducing a PTC that targets the mutant mRNA for degradation by NMD (Popp and Maquat 2013).

Pseudo-exon inclusion was first reported in β-Thalassemia patients (Dobkin, Pergolizzi et al. 1983, Treisman, Orkin et al. 1983). A T>G mutation located 705 bp downstream of the *HBB* middle exon created a new donor splice site and activated an acceptor splice site present within the second intron (Dobkin, Pergolizzi et al. 1983). Several additional deep intronic mutations leading to pseudo-exon inclusion have since been identified in patients affected by multiple disorders (Vaz-Drago, Custodio et al. 2017).

The longer a gene the more likely it is to be affected by pathogenic mutations (Lopez-Bigas, Audit et al. 2005). It is therefore not surprising that numerous deep intronic mutations have been described in particularly long genes such as those associated with neurofibromatosis (Cunha, Oliveira et al. 2016) and Duchenne muscular dystrophy (Beroud, Carrie et al. 2004, Gurvich, Tuohy et al. 2008, Trabelsi, Beugnet et al. 2014, Gonorazky, Liang et al. 2016). Remarkably, deep-intronic mutations that promote inclusion of a pseudo-exon have been described in several hereditary tumor syndromes (Vaz-Drago, Custodio et al. 2017). A splicing enhancer created *de novo* within an intronic region may be sufficient to promote recognition by the spliceosome leading to pseudo-exon inclusion.

Most deep intronic mutations have no effect on canonical splice sites. Yet, some mutations that create a new splice site interfere with recognition of natural splice sites, resulting in the inclusion of intronic sequences in the mRNA that introduce a PTC, triggering the mutant mRNA for degradation by NMD.

Besides creating new exon-intron boundaries, deep-intronic mutations can also inactivate intron-encoded RNA genes. Indeed, point mutations in the *RNU4ATAC* gene were identified in patients affected by the developmental disorder Taybi-Linder syndrome (TALS) or Microcephalic ostedysplastic primordial dwarfism type 1 (MOPD1) and Roifman syndrome (Edery, Marcaillou et al. 2011, He, Liyanarachchi et al. 2011, Merico, Roifman et al. 2015).

Consistent with loss-of-function of the mutant snRNA, higher levels of unspliced U12-type introns were detected in patient-derived fibroblasts.

Moreover, multiple cases of genetic diseases caused by deep intronic mutations that disrupt transcription regulatory motifs have been identified, resulting in the alteration of the transcription of the mutant gene. For example, the first intron of the MPZ gene contains binding sites for transcription factors SOX10 and EGR2, which are implicated in the regulation of MPZ expression (Antonellis, Dennis et al. 2010). The MPZ protein is required for proper myelination and several coding and splice site mutations in the MPZ gene have been described as cause of Charcot-Marie-Tooth disease type 1B, a demyelinating peripheral neuropathy.

Disease-causing mutations at 3' end of transcripts

As in splicing, the use of alternative polyadenylation sites contributes to the complexity of the transcriptome and thereby affects important cellular functions in physiological as well as pathophysiological conditions. Interestingly, the spliceosomal component U1A has been identified as a key player in the regulation of 3' end processing of the *SMN* pre-mRNA. SMN protein is involved in small nuclear ribonucleoprotein (snRNP) biogenesis and its insufficient expression causes spinal muscular atrophy (SMA). Binding of U1A to the *SMN* 3'-UTR specifically inhibits cleavage of the pre-mRNA, causing a decrease in polyadenylation and a corresponding decrease in SMN protein expression (Workman, Veith et al. 2014).

As in SMA, disruption of 3' end processing represent a common feature of other neurological, oncological, immunological and haematological disorders (Curinha, Oliveira Braz et al. 2014).

Disease-associated 3' end aberrant polyadenylation profiles can be caused by mutations that disrupt or introduce new pA signals and deregulation of expression of cleavage or polyadenylation factors (**Figure 7A, E**). These mutations lead to shortening/lengthening of 3' UTR due to differential usage of upstream/downstream (proximal/distal) pA sites, or read-through (transcription interference) due to the complete unrecognition of the poly(A)

signal site, leading the RNAPII to continue transcribing uninterrupted into the intergenic region and downstream gene (Danckwardt, Hentze et al. 2008).

Early work on mutations in pA signals was performed in human β-globin and α_2-globin genes isolated from patients with β- and α-thalassemias (Higgs, Goodbourn et al. 1983, Orkin, Cheng et al. 1985). The β-globin pA mutation causes the activation of a distal pA site, which lead to lengthening of 3' UTR and decrease in β-globin mRNA expression. In the case of α_2-globin, it was found that pA mutations affect RNAPII transcription termination and expression of the downstream gene α1-globin. Misregulation of both 3' end processing and transcription termination has also been associated with other human diseases. Indeed, cancer cells present transcriptional read-through profiles generated by the inactivation of pre-mRNA cleavage and transcription termination. Pervasive transcription most often leads to transcription interference of the neighbouring gene or read-through fusion transcripts that result from *cis*-splicing of one pre-mRNA molecule synthesized from two adjacent genes in the same coding orientation (Kumar-Sinha, Kalyana-Sundaram et al. 2012, Varley, Gertz et al. 2014, Grosso, Leite et al. 2015, He, Yuan et al. 2018). On the other hand, widespread increase in the usage of proximal pA signals of onco-related mRNAs has also been observed in various cancer types. Shortening of the 3' UTR increases stability and translation of mutant mRNA, mostly by repressing miRNA-mediated degradation and RNPs regulation (Sandberg, Neilson et al. 2008). Yet, the extent to which the global reported alterations at the 3' end represent the cause or the consequence of upstream deregulated processes remains to be clarified.

Another example of defeated transcription termination regulated at the chromatin level was observed in cells undergoing senescence. Transcriptome profiling of non-proliferating cells led to the identification of a new family of antisense RNAs, named START RNAs, that were produced during cellular senescence by transcriptional read-through at convergent protein-coding genes (Muniz, Deb et al. 2017).

Furthermore, infection of human cells by influenza or herpes simplex virus causes widespread misregulation of 3' end processing and transcription termination (Noah, Twu et al. 2003, Rutkowski, Erhard et al. 2015).

Interestingly, a new class of read-through transcripts, produced upon osmotic and oxidative stress or heat shock, was recently found in mammalian cells and termed downstream of gene-containing transcripts (DoGs) (Vilborg, Sabath et al. 2017).

The 3' end processing machinery includes more than 50 proteins (Shi, Di Giammartino et al. 2009) and deregulation of expression of such proteins has been associated to human disease (Curinha, Oliveira Braz et al. 2014). One such example is the autosomal dominant oculopharyngeal muscular dystrophy that is caused by $(GCG)_{8-13}$ expansions in the coding region of the *PABPN1* gene. This expansion results in an increase of self-association and misfolding of the PABPN1 protein in skeletal muscle (Banerjee, Apponi et al. 2013). It has been shown that mutated PABPN1 sequesters polyadenylated mRNAs in nuclear inclusion bodies (Calado, Tome et al. 2000) and that its downregulation leads to an increase in the recognition of proximal pA sites (Jenal, Elkon et al. 2012).

Correct pre-mRNA processing is crucial for the export of mRNAs to the cytoplasm, their translation efficiency and approval by quality control mechanisms.

1.5. Measuring mRNA biogenesis in health and disease

To understand the cell and molecular biology of genetic diseases it is essential to study the mechanisms that actively participate in mRNA biogenesis, such as transcription and pre-mRNA processing. Measurement of the reaction rates of transcription, splicing and 3' end processing can reveal defects in gene expression and highlight the quality control checkpoints that may be active in disease.

Since transcription occurs in a specific location in the cell and processing mechanisms are mostly co-transcriptional, the selection of the cellular models and assays to study mRNA biogenesis in health and disease is critical to accomplish a physiologically relevant result. Together with induced pluripotent stem cells (iPSCs) derived from patients, patient-derived lymphoblastoid cell lines (LCLs) are recognized as relevant disease models to study gene regulation (Soldner and Jaenisch 2012, Kumar, Curran et al. 2016). However, compared to LCLs, well-stablished cell lines as human embryonic kidney 293 (HEK 293) cells are easier to culture and transfect, allowing further dissection of the molecular mechanisms and players that are disrupted in disease. Moreover, well-stablished cell lines also provide a pure population of cells, which is valuable since it provides a consistent sample and reproducible results.

Numerous approaches can be used to determine the precise kinetics and timing of mRNA biogenesis. This diverse range of methodologies can be divided in two main types: biochemical-based and microscopy-based (**Figure 9**). Both methodologies have certain advantages over the other and are used for different purposes, for example, biochemical assays allow the purification of specific RNA populations, while confocal microscopy allow single-cell and single-molecule analysis of mRNA biogenesis. The integration of information coming from both methodologies provides a more complete insight into how, when and where different mRNA biogenesis processes contribute to regulation of gene expression in health and disease.

In the 1960s, it was found that peripheral mature B lymphocytes infected with Epstein-Barr virus (EBV) acquired the ability to grow continuously in culture (Miller 1982, Sie, Loong et al. 2009). Since then, EBV-transformed B cells have been extensively used in biomedical research studies. Lymphocytes are mature cells derived from differentiation of immature lymphoblast cells. Infection by the EBV induces transformation of the mature and fully differentiated cells into lymphoblastoid cells.

EBV is one of eight human herpesviruses and it is carried by over 90% of the world's adult population (Scott 2017). Its double-stranded DNA genome can either be circular or linear, depending if it is in the latent or lytic stage of the life cycle, respectively. The establishment of a lymphoblastoid cell line involves the isolation of peripheral blood lymphocytes that includes B, T and Natural Killer cells, infection with EBV followed by transformation of mature B cells, expansion and cryopreservation of the established lymphoblastoid cell line (LCL) (Tosato and Cohen 2007). Upon transformation, the circular viral genome is maintained as an extrachromosomal element with a chromatin structure that resembles the host chromatin. Lymphoblastoid cells serve as the EBV latency compartment where silencing of viral gene expression allows maintenance of the viral genome, avoidance of immune surveillance, and life-long carriage. EBV encodes a number of viral factors that interact with chromatin and chromatin remodelers of the lymphoblastoid cells (Scott 2017). One of the best known viral factors is EBNA2 that interacts with sequence specific DNA binding protein to regulate EBV latency gene expression in B cells and to modify cellular gene expression which results in stimulation of G0 to G1 cell cycle progression resulting in B-cell “immortalization” (Zhou, Schmidt et al. 2015). Interestingly, infected B cells can support the lytic stage of the virus life cycle, but they more frequently host nonproductive infections through expression of a limited number of latent EBV genes (latency III genes) that drive proliferation of B cells as an alternative mechanism of expanding the infected cell pool (Williams, Quinn et al. 2015).

Once established, LCLs are commonly considered as “immortal” cells, however they exist in two distinct cellular stages: preimmortal and postimmortal (Sie,

Loong et al. 2009, Hussain and Mulherkar 2012). Preimmortal LCLs have diploid karyotypes and normal telomerase activity. After 160 to 180 population doublings in cell culture, most preimmortal LCL start dying due to shortening of telomeres. The EBV-transformed cells that survive develop different tumorigenic proprieties, namely strong telomerase activity, aneuploidy, and deregulation of gene expression (Sugimoto, Tahara et al. 2004, Jha, Pei et al. 2016). In the preimmortal stage, LCLs can be used to perform a variety of genetic and functional studies, overcoming the need of resample individuals.

There are a number of cell repositories that provide essential source of lymphoblastoid, primarily fibroblasts cell lines and, more recently, iPSC lines. This constitutes a source of patient's DNA, RNA and protein, which can then be used to study the molecular pathways disrupted in diseased cells (Consortium 2012).

The Human embryonic kidney 293 (HEK 293) cell line was derived in 1973 by transformation of primary embryonic kidney cell culture with sheared DNA of adenovirus type 5 (Graham, Smiley et al. 1977) and has been the most frequently used after HeLa in cell biological studies. However, its origin, phenotype and karyotype are still a concern (Hughes, Marshall et al. 2007). HEK 293 cells are considered kidney epithelial cells, however HEK 293 cells do not demonstrate tissue-specific gene signature, expressing fibroblastic, endothelial and epithelial markers. HEK 293 cells are considered female referring to the presence of several X chromosomes and lack of Y chromosome. Moreover, HEK 293 cells have an aberrant karyotype with a chromosome number ranging between 62 and 76, depending on the cell banks or laboratories (Stepanenko and Dmitrenko 2015).

Despite that, HEK 293 cells have been widely used since they are easy to grow in culture and to transfect with commercially available kits. Thus, HEK 293 is an excellent cell line to use in transfection experiments or to produce recombinant DNA or gene products and an ideal cell line for therapeutic protein and virus production by the biotechnology industry [reviewed in (Bussow 2015)].

Changes in the total amount of a certain mRNA are closely coordinated with transcriptional, processing and decay rates and together these events have profound effects on gene expression during development and disease [reviewed in (Bentley 2014, Custodio and Carmo-Fonseca 2016, Zinder and Lima 2017)]. Thus, analysis of total RNA levels does not match changes in transcription rates, but are inherently dependent on the RNA half-life of the respective transcript. To study early processes in gene expression it is crucial to analyse nascent transcripts. This has been achieved by isolating chromatin-associated RNA (Wuarin and Schibler 1994, Bhatt, Pandya-Jones et al. 2012) or immunoprecipitation of chromatin-associated RNAPII-bound transcripts (Mayer, di Iulio et al. 2015, Nojima, Gomes et al. 2015). Additional methodologies to investigate newly synthesized RNA molecules involve metabolic labelling of nascent transcripts (Fuchs, Voichek et al. 2015) and include global nuclear run on sequencing (GRO-seq) (Core, Waterfall et al. 2008), precision nuclear run on sequencing (PRO-seq) (Kwak, Fuda et al. 2013), 4sUDRB-seq (Fuchs, Voichek et al. 2014) and transient transcriptome sequencing (TT-seq) (Schwalb, Michel et al. 2016).

Metabolic labelling combined with chemical modification and fractionation of labelled RNAs has allowed the isolation of nascent transcripts and the subsequent calculation of the rate of production of a certain pre-mRNA. In this technique, newly-transcribed RNAs are labelled *in vivo* by incorporation of thio-substituted uridines, which are subsequently biotinylated *in vitro* and purified on streptavidin-coated magnetic beads (Cleary, Meiering et al. 2005, Dolken, Ruzsics et al. 2008). Upon addition to the culture medium, the naturally occurring nucleoside analog 4-thiouridine (4sU) is rapidly taken up by mammalian cells, integrated into the cell's ribonucleotide pool via the salvage pathway (Lane and Fan 2015) and incorporated into the growing RNA chain in place of endogenous uridine (Melvin, Milne et al. 1978).

Newly transcribed RNA depicts the transcriptional activity of every gene during the timeframe of 4sU exposure. 4sU-tagging in the timescale of minutes thus provides information about transcription rate for a certain gene. Combination of reversible inhibition of transcription elongation and 4sU-tagging found that

most genes are transcribed at about 3.5 kb/min, with elongation rates varying between 2 kb/min and 6 kb/min (Fuchs, Voichek et al. 2015).

Contrasting with a plethora of studies that either measure global mRNA levels or focus on transcription, technical challenges have limited our understanding of RNA processing dynamics. A diverse range of methods has been used to measure the kinetics of human precursor messenger RNA (pre-mRNA) splicing. The kinetic coupling between transcription and splicing, impose a time constraint on splicing to take place. Metabolic RNA labelling approaches have been applied in human (Windhager, Bonfert et al. 2012), mouse (Rabani, Raychowdhury et al. 2014) and yeast (Barrass, Reid et al. 2015) cells to study the kinetics of pre-mRNA splicing genome-wide. The resulting estimates of the time necessary to excise human introns show a broad variation between 30 seconds (Huranova, Ivani et al. 2010, Martin, Rino et al. 2013) and 2-5 minutes (Schmidt, Basyuk et al. 2011, Windhager, Bonfert et al. 2012, Coulon, Ferguson et al. 2014). These discrepant results may be related to the kinetic information that different methods provide.

In conclusion, this approach allows for the direct analysis of the dynamics of RNA synthesis and processing in mammalian cultured cells.

Microscopy-based analysis

Much of knowledge about the fundamentals of mRNA biogenesis come from ensemble *in vitro* biochemical approaches. However, using conventional biochemical techniques only average behaviour of cells and molecules can be measured. On the other hand, imaging allows *in vivo* and single-cell measurements along with the ability to measure the behaviour of individual molecules in time and space and the distribution of signal around the mean (Larson, Singer et al. 2009, Coulon, Chow et al. 2013).

In vivo single-molecule RNA imaging approaches have been extensively used to study processes involved in mRNA biogenesis and localization (Brody, Neufeld et al. 2011, Schmidt, Basyuk et al. 2011, Martin, Rino et al. 2013, Yoon, Wu et al. 2016, Haimovich, Ecker et al. 2017). Studies utilizing live cell imaging for the analysis of transcription and splicing kinetics rely primarily on

reporter genes in human cells. The approach is based on the fusion of the bacteriophage MS2/PP7/λN coat protein to a fluorescent protein and a reporter mRNA containing multiple RNA stem-loops that are recognized by the MS2 coat protein. Binding of MS2-GFP fused protein to the RNA stem-loops allows directly visualization of single RNAs made from single genes by confocal microscopy. Depending on the localization of the stem-loops within the transcript body, it is possible to measure transcription, pre-mRNA processing kinetics and coupling between those processes. For example, three studies used stably integrated β-globin reporter genes with MS2 or PP7 stem loops inserted into intronic or exonic sequences to track pre-mRNA. One of such studies was performed in HEK 293 cells and used stably integrated β-globin reporter gene with MS2 or PP7 stem loops inserted into intronic sequences and reported that intron removal takes around 20 seconds after transcription (Martin, Rino, et al 2013). Other live-cell studies analysed an ensemble population of β-globin pre-mRNAs (Coulon, Ferguson et al. 2014) or adenovirus derived reporter pre-mRNAs (Schmidt, Basyuk et al. 2011) and concluded that splicing takes about 5 minutes for completion. Notably, the single-molecule approach revealed that transcripts exhibit stochastic processing, with some transcripts being spliced co-transcriptionally and others spliced post-transcriptionally. These discrepant results may be related to the fact that when multiple nascent RNAs are simultaneously detected a modelling approach must be applied to infer kinetic information, whereas direct analysis of individual pre-mRNA molecules provides a dynamic level of information that is not possible to obtain in ensemble measurements.

Some of these processes have been examined in great depth, while others have not. For example, there are only a few studies of transcription termination in eukaryotes and none of them used single-molecule tracking of pre-mRNA (Boireau, Maiuri et al. 2007, Coulon, Ferguson et al. 2014) .

A quantitative understanding of regulation of transcription and processing in subpopulations of cells or single cells would shed light on identifying *cis*-regulatory motifs or *trans*-acting factors that may be important for disease progression and development.

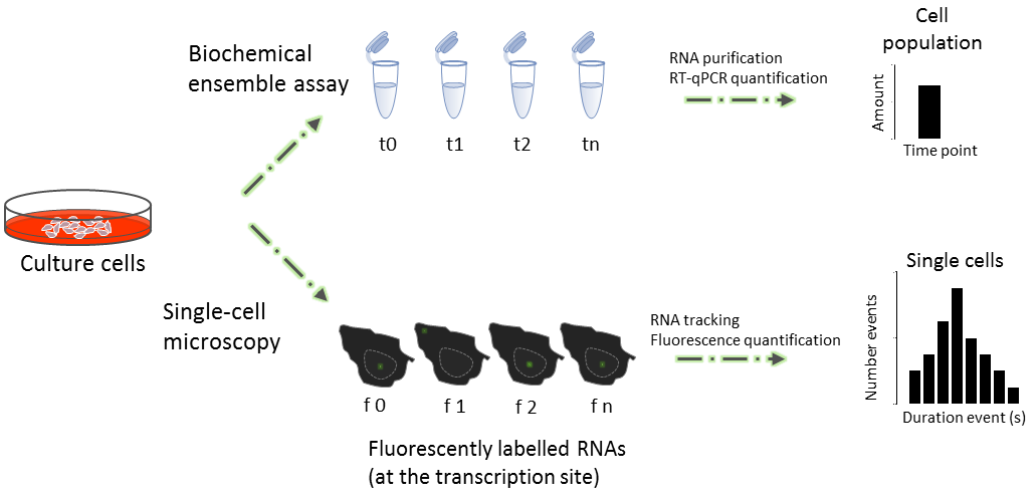


Figure 9 – Ensemble versus single-cell assays.

2. Objectives

For my PhD training, I proposed to explore the impact of the deregulation of mRNA biogenesis in the context of human genetic disease. My work focused on four main hypotheses:

Hypothesis I: A checkpoint for mRNA quality operates co-transcriptionally, before NMD, reducing the production of potentially deleterious proteins encoded by genes altered by disease-causing splice-site mutations.

To address this hypothesis, mRNA kinetics in terms of its production rate, release and transport to the cytoplasm was analysed in cell lines derived from patients with genetic diseases caused by mutations that affect splicing.

Hypothesis II: DNA variants located throughout intronic regions are an important cause of human genetic diseases.

To pursue this idea I reviewed evidence from mRNA analysis and genomic sequencing indicating that pathogenic mutations can occur deep within the introns.

Hypothesis III: Different biochemical methodologies may introduce bias that lead to the overrepresentation of a long transcripts, which in turn may interfere with the calculation of splicing efficiencies.

To test this hypothesis, I compared splicing efficiencies of different protein-coding nascent RNAs purified using three different biochemical methods.

Hypothesis IV: The kinetics of the 3' end processing of pre-mRNA molecules is highly regulated and can be measured by live-cell microscopy with single-molecule resolution.

To test this hypothesis, I measured time of residence of single pre-mRNA molecules at transcription sites of two different transgenes with MS2 or PP7 binding sites in the respective last exon.

3. Results

3.1. Transcription-coupled RNA surveillance in human genetic diseases caused by splice site mutations

The results presented below were published in the *Human Molecular Genetics* peer-reviewed journal. The article in the publication format can be found in the Appendix of this thesis.

Rita Vaz-Drago, Marco T. Pinheiro §, Sandra Martins, Francisco J. Enguita, Maria Carmo-Fonseca*, and Noélia Custódio*

Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, 1649-028 Lisboa, Portugal

§ *Present address: Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, United Kingdom.*

Author contribution

Rita Vaz-Drago, Noélia Custódio and Maria Carmo-Fonseca designed the experiments and wrote the manuscript. Rita Vaz-Drago performed most of the experiments and data analysis. Marco T. Pinheiro performed sub-cellular fractionation of wild-type and *TAZ* cell lines. Sandra Martins contributed with expertise and helped in Western Blot experiments. Francisco J. Enguita analysed a microarray dataset. All authors revised part or the entire manuscript.

3.1.1. Overview

The work described in this chapter addresses the significance of a co-transcriptional RNA quality control mechanism in the context of human genetic diseases caused by mutations in canonical splice sites.

Current estimates indicate that approximately one third of all disease-causing mutations are expected to disrupt splicing. Abnormal splicing often leads to disruption of the reading frame with introduction of a premature termination codon that targets the mRNA for degradation in the cytoplasm by nonsense mediated decay (NMD). In addition to NMD, there are surveillance mechanisms that act in the nucleus, while transcripts are still associated with the chromatin template. However, those nuclear quality control mechanisms have been mainly described in yeast. Here we used patient-derived lymphoblastoid cell lines as disease models to address how biogenesis of mRNAs is affected by mutations located within canonical splice sites. Most of the mutations that disrupt canonical splice sites lead to exon-skipping or to the activation of a nearby cryptic splice site, leading to the introduction of a premature termination codon. As consequence, these mutant mRNAs are triggered to mRNA degradation in the cytoplasm. However, for some mutant transcripts, RNA levels associated with chromatin were found down-regulated. Quantification of nascent transcripts further revealed that a subset of genes containing splicing mutations have reduced transcriptional activity. Following treatment with the translation inhibitor cycloheximide the cytoplasmic levels of mutant RNAs increased, while the levels of chromatin-associated transcripts remained unaltered. These results suggest that transcription-coupled surveillance mechanisms operate independently from NMD to reduce cellular levels of abnormal RNAs caused by mutations located within canonical splice sites.

3.1.2. Transcripts with splicing mutations are less abundant in the nucleoplasm of patient-derived cells

Co-transcriptional RNA quality control mechanisms play a role in reducing the levels of RNA molecules with defects in processing. As most of these studies have been conducted using yeast genes and human reporter genes as models, we asked whether co-transcriptional RNA quality control plays a physiological role in the context of human genetic diseases. To address this issue, we used patient-derived lymphoblastoid cell lines to analyse the life cycle of RNAs produced from genes that contain naturally-occurring disease-causing splicing mutations. To obtain quantitative information on RNA levels, we used a biochemical fractionation approach that allows dynamic changes in transcription or nuclear RNA degradation to be distinguished from changes in cytoplasmic steady-state mRNA levels (**Figure 1A**). We optimized for lymphoblastoid cells a fractionation technique that was initially described by Wuarin and Schibler (Wuarin and Schibler 1994) and subsequently modified in the Proudfoot and Black laboratories (Dye, Gromak et al. 2006, Pandya-Jones and Black 2009). The protocol takes advantage of the fact that once RNA polymerase II (RNAPII) initiates transcription it forms a tight complex with the DNA template that resists treatment with urea and mild detergent. The extraction procedure does not dissociate histones from DNA and therefore the chromatin remains highly compacted and can be sedimented with associated nascent transcripts by low-speed centrifugation. Transcripts detected in the nucleoplasmic supernatant fraction are assumed to have been released from the DNA template. The efficiency of the fractionation protocol was assessed by western blotting (**Figure 1B**) and RT-PCR (**Figure 1C**). For the western blotting assay antibodies against lamin A/C, β-actin, U2B'' and histone H3 proteins were used (**Figure 1B**). Actin was found predominantly in the cytoplasmic fraction, whereas U2 snRNP specific protein B'' (U2B''), lamin A/C and histone H3 were detected exclusively in nuclear fractions. The nucleoplasmic fraction should contain nuclear proteins that either do not associate with chromatin or are loosely attached to chromatin. The U2B'' is mostly detected in the nucleoplasmic fraction as previously reported for other components of the spliceosome (Pandya-Jones and Black 2009). Lamin and histone H3 are well known chromatin-associated proteins and, accordingly, they are

predominantly detected in the chromatin fraction. To characterize the RNA species present in each fraction, RT-PCR analysis was carried out with primers for total, unspliced and spliced Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) mRNA, as well as primers for Xist RNA (**Figure 1C**). The results clearly show that unspliced GAPDH pre-mRNA is restricted to the nucleus and localizes predominantly in the chromatin fraction. Spliced mRNA is also detected in the chromatin fraction, consistent with the view that most splicing occurs co-transcriptionally (Bentley 2014), but is most abundant in the cytoplasm. The distribution of total GAPDH RNA is similar to that of spliced mRNA, as expected considering that mature transcripts are transported and accumulate in the cytoplasm. A completely different distribution pattern is observed for Xist RNA, which is restricted to the nucleus and predominantly localized in the chromatin fraction consistent with its well established physical interaction with the X-chromosome (Engreitz, Pandya-Jones et al. 2013).

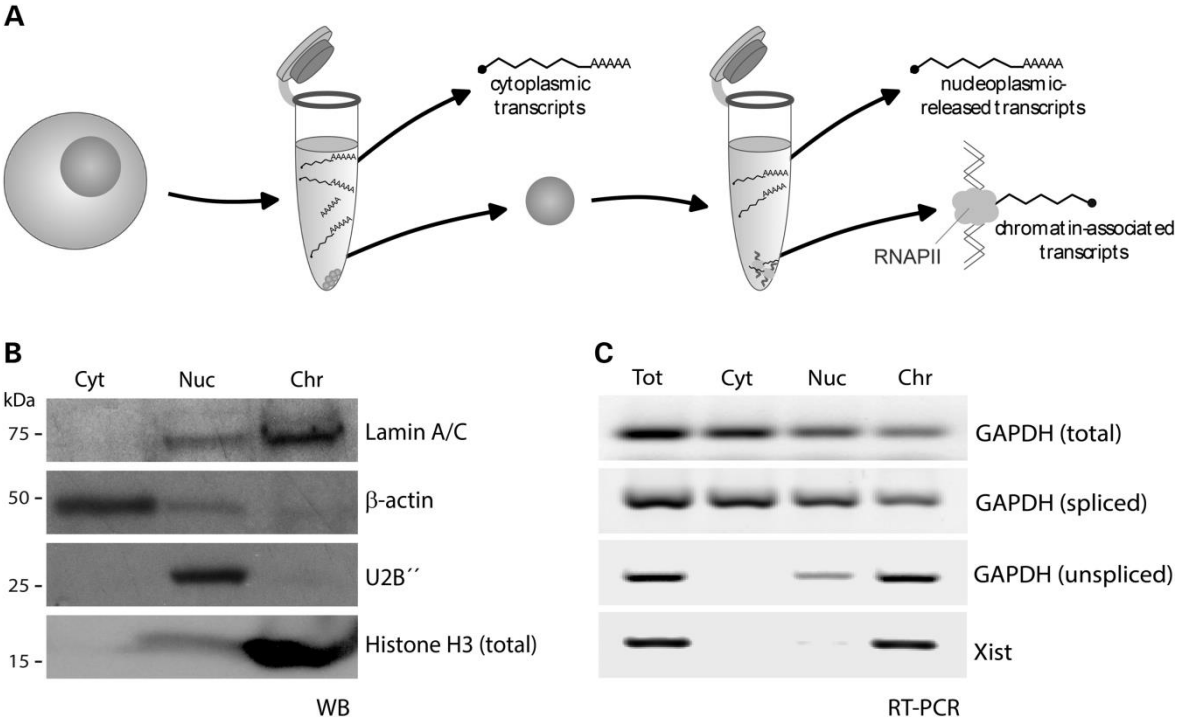


Figure 1. Sub-cellular fractionation.

A) Illustration of the sub-cellular fractionation procedure. After cell lysis, nuclei are separated from the cytoplasmic (cyt) fraction by centrifugation. Nuclei are then treated with urea and nonionic detergent. Upon centrifugation, the chromatin-associated fraction (chr) sediments separating from the soluble nucleoplasmic fraction (nuc). B) Western Blotting analysis. Lymphoblastoid cells (GM16113) were fractionated and analysed by Western Blotting (WB) for detection of Lamin A/C, β-actin, U2B" and Histone H3 (total). Equal amounts of total protein from each fraction were loaded per

lane. C) RT-PCR analysis. RNA was isolated from lymphoblastoid cells (GM04490), reverse transcribed with random primers and PCR amplified using primers for total, spliced and unspliced GAPDH RNA and total Xist RNA. Equal amounts of cDNA from total and fractionated samples were loaded per lane.

Having validated the fractionation methodology, we next determined RNA levels in cell lines derived from a healthy donor and from patients affected by three distinct monogenic recessive disorders associated with splicing mutations: Barth syndrome (OMIM 302060), Deafness, autosomal recessive 49 (OMIM 610153) and Xeroderma Pigmentosum (OMIM 278720).

Barth syndrome is an X-linked recessive syndrome caused by mutations in the *TAZ* gene (MIM: 300394), which codes for an acyltransferase required for remodeling of cardiolipin in the inner mitochondrial membrane. TAZ loss of function results in an inborn error of lipid metabolism (Bione, D'Adamo et al. 1996, Gonzalez 2005, Kirwin, Manolakos et al. 2014). We analysed two patient-derived cell lines, each containing a point mutation that affects splicing of the *TAZ* gene. The splicing mutations (SM) localize in intron 1, at the 5' and 3' splice sites (**Table 1 and Figure 2A**). It was previously shown that the 5' splice site mutation activates two cryptic donor splice sites either upstream or downstream of the point mutation, and the 3' splice site mutation can either activate a cryptic acceptor splice site within exon 2 or lead to exon 2 skipping (Johnston, Kelley et al. 1997). Most 5'SM transcripts expressed in lymphoblastoid cells correspond to the longer splice product, which does not disrupt the open reading frame. The less abundant shorter splice product has the open reading frame disrupted (Johnston, Kelley et al. 1997) The two splice products resulting from the 3' splice site mutation are expressed at similar levels and only one has the open reading frame disrupted (Johnston, Kelley et al. 1997) For comparison, we analysed a cell line with a point mutation in exon 2 that introduces a PTC without affecting the splicing frame (Gonzalez 2005).

Cell Line	Reference	Affected Gene	Mutation	Transcript
GM16113		-	WT	WT
GM22129	Patient 2 (Johnston, Kelley et al. 1997)	*TAZ*, Xq28	IVS1+5G>A (5'SM)	Exon 1 cryptic, PTC Intron 1 cryptic
GM22165	Patient 4 (Johnston, Kelley et al. 1997)		IVS1-2A>G (3'SM)	Exon 2 cryptic, PTC Exon 2 skipped
GM22150	Patient 2 (Gonzalez 2005)		Trp79Ter (PTC)	PTC
GM20190	PKDF399 (Riazuddin, Ahmed et al. 2006)	*MARVELD2*, 5q13.2	IVS3-1G>A (3'SM)	Exon 4 cryptic, PTC
GM20193	PKDF068 (Riazuddin, Ahmed et al. 2006)		IVS4+2T>C (5'SM_1)	Intron 4 cryptics, PTC
GM20172	PKDF443 (Riazuddin, Ahmed et al. 2006)		IVS4+2delTGAG (5'SM_2)	Intron 4 cryptics, PTC
GM20189	PKDF340 (Riazuddin, Ahmed et al. 2006)		Arg500Ter (PTC)	PTC
GM04490	XP25BE (Khan, Oh et al. 2006)	*XPC*, 3p25.1	IVS11-1_IVS11-2 delAG IVS11-6_IVS11-7 InsCC (3'SM)	Intron 11 retained, PTC Exon 12 cryptic, PTC Exon 12 skipped, PTC

Table 1. Cell lines used in this study.

RNA levels were measured by quantitative real-time PCR (RT-qPCR) using the primers indicated in **Figure 2A** **Figure 2B** depicts the level of mutant transcripts in total cellular RNA as fold change relative to values detected using the same primer sets in cells from a healthy donor. The abundance of each PCR product was normalized to the level of GAPDH RNA detected in the same RT-qPCR run. The results show that the three mutant transcripts analysed are significantly less abundant than wild-type *TAZ* RNA (**Figure 2B**), in agreement with the loss of function phenotype observed in patients. Next we determined the levels of

mutant transcripts in the cytoplasm, nucleoplasm and chromatin, using equal amounts of RNA from each fraction. The cytoplasmic levels of the three mutant *TAZ* RNAs are significantly reduced relative to the wild-type (**Figure 2C**). However, distinct scenarios are observed in nuclear fractions (**Figure 2D, E**). Mutant transcripts that contain a PTC but have normal splicing do not significantly differ from wild-type, suggesting that these RNAs are exclusively degraded in the cytoplasm (**Figure 2D and E, PTC**). In contrast, transcripts with the 5' splice site mutation are significantly less abundant than wild-type transcripts in both nucleoplasm and chromatin fractions (**Figure 2D and E, 5'SM**). The level of transcripts with the 3' splice site mutation is similar to wild-type in the nucleoplasm (**Figure 2D, 3'SM**), but higher in the chromatin (**Figure 2E, 3'SM**). This heterogeneity of results prompted us to analyse additional mutant transcripts associated with unrelated diseases.

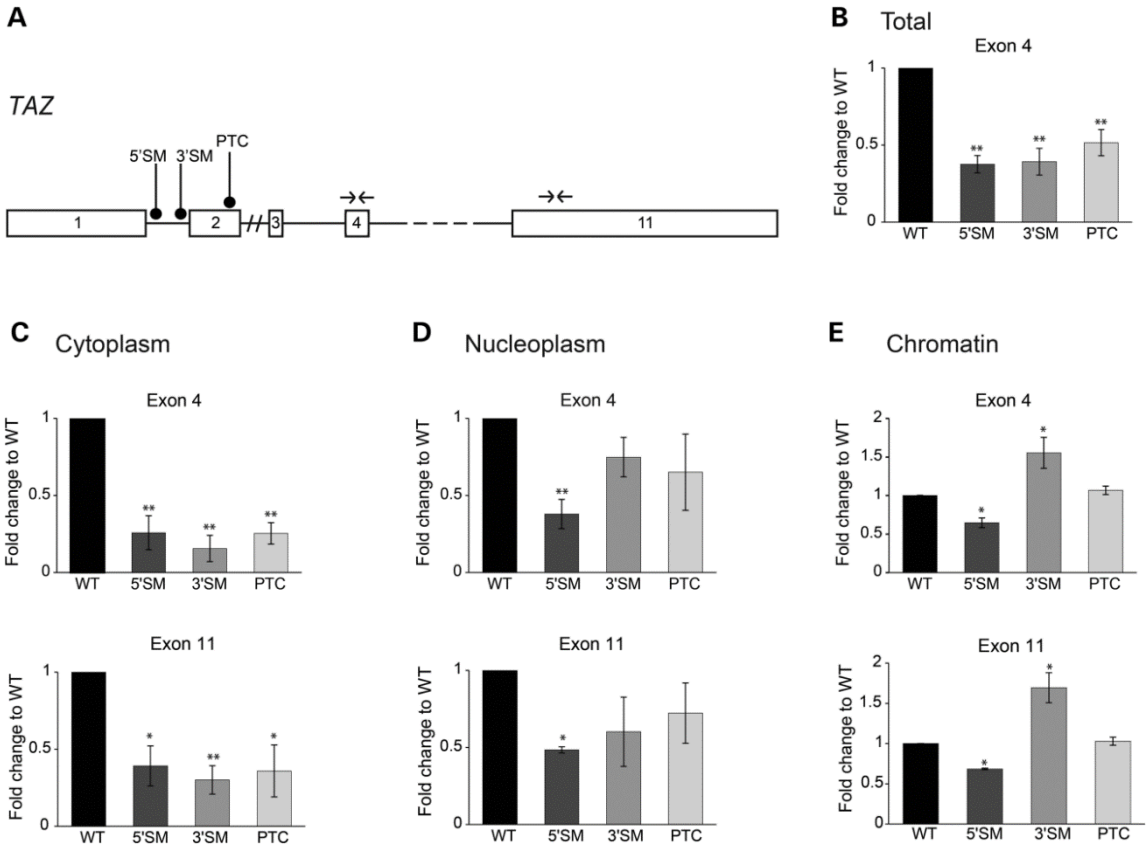


Figure 2. Sub-cellular distribution of wild-type and mutant *TAZ* transcripts.
A) Illustration of the *TAZ* gene structure (total length: 10185 bp). Exons are represented by numbered boxes and introns by lines; doubled intersected lines denote introns with more than 1000 bp. Gene region from exon 5 to exon 10 is represented

by a dashed line. Positioning of mutations (5'SM, 3'SM, PTC) and primers used for PCR amplification (paired arrows) are indicated. B) Total cellular RNA was extracted from the indicated cell lines, reverse transcribed with random primers and analysed by RT-qPCR using primers for exon 4. The amount of PCR product obtained from each cell line was normalized to the level of GAPDH RNA detected in the same line. C) RNA was extracted from sub-cellular fractions isolated from each cell line and analysed by RT-qPCR using primer sets for exon 4 and exon 11. The amount of PCR product obtained from each fraction was normalized to the level of GAPDH RNA detected in the same fraction. In all graphs shown, data are expressed as fold change relative to the levels of wild-type (WT) transcripts. The histograms depict mean and standard deviation of three independent experiments. The asterisk denotes statistically significant differences (Student's *t*-test, $*p<0.05$, $**p<0.01$).

Deafness, autosomal recessive 49 is a congenital profound sensorineural hearing loss of all frequencies, caused by dysfunction of a tricellulin protein coded by the *MARVELD2* gene (MIM: 610572). Tricellulin is a tight-junction protein that contributes to the structure and function of tricellular contacts of neighboring cells. Loss of function of this protein may selectively affect the cellular permeability to ions or small molecules, resulting in a toxic microenvironment for cochlear hair cells and subsequently ear loss (Riazuddin, Ahmed et al. 2006, Nayak, Lee et al. 2013). We analysed cell lines derived from three patients, each homozygous for a distinct splice site mutation in the *MARVELD2* gene (**Table 1 and Figure. 3A**). The splice site mutations localize in intron 3 at the 3' splice site, and in intron 4 at the 5' splice site. The 5' splice site mutations activate cryptic donor sites in intron 4, and the 3' splice site mutation activates a cryptic acceptor site within exon 4; all the mutations lead to the production of mRNAs containing PTCs due to shifts in the open reading frame (Riazuddin, Ahmed et al. 2006). For comparison, we analysed a cell line homozygous a point mutation in exon 5 that introduces a PTC without affecting splicing (Riazuddin, Ahmed et al. 2006).

RNA levels were measured by RT-qPCR using the primers indicated in **Figure 3A**. Similarly to the results obtained with *TAZ* transcripts, the total cellular levels of the four mutant *MARVELD2* RNAs are significantly reduced compared to wild-type (**Figure 3B**). Analysis of RNA levels in sub-cellular fractions reveals that mutant transcripts are significantly less abundant in the cytoplasm (**Figure 3C**), in agreement with the finding that they all contain PTCs. In nuclear

fractions the levels of mutant transcripts that contain a PTC but have normal splicing are similar to wild-type (**Figure 3D and E, PTC**), indicating that these RNAs are exclusively degraded in the cytoplasm. However, all transcripts with splicing mutations are significantly less abundant in the nucleoplasm (**Figure 3D**). In the chromatin fraction, significantly reduced levels are only detected for the 3' splice site mutant only (**Figure 3E, 3'SM**).

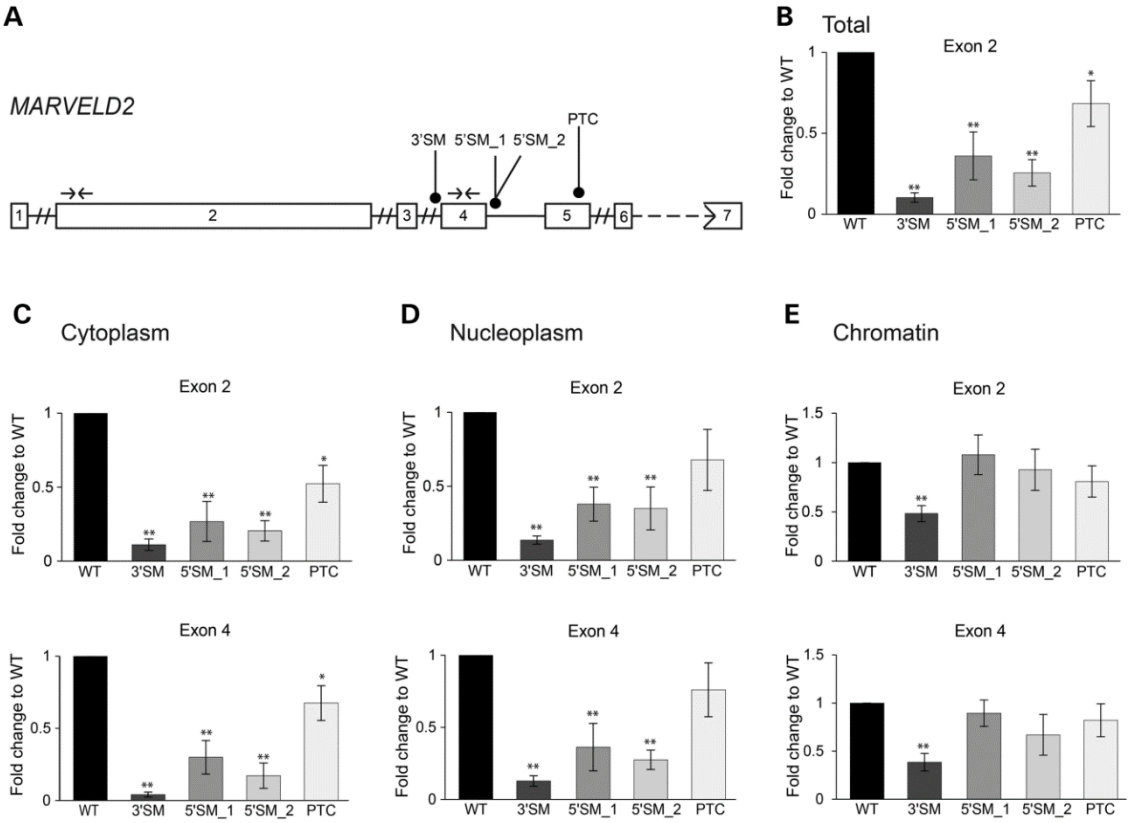


Figure 3. Sub-cellular distribution of wild-type and mutant *MARVELD2* transcripts.
A) Illustration of the *MARVELD2* gene structure (total length: 27762 bp). Exons are represented by numbered boxes and introns by lines; doubled intersected lines denote introns with more than 1000 bp. Intron 6 is represented by a dashed line. Positioning of mutations (3'SM, 5'SM_1, 5'SM_2, PTC) and primers used for PCR amplification (paired arrows) are indicated. B) Total cellular RNA was extracted from the indicated cell lines, reverse transcribed with random primers and analysed by RT-qPCR using primers for exon 2. The amount of PCR product obtained from each cell line was normalized to the level of GAPDH RNA detected in the same line. C) RNA was extracted from sub-cellular fractions isolated from each cell line and analysed by RT-qPCR using primer sets for exon 2 and exon 4. The amount of PCR product obtained from each fraction was normalized to the level of GAPDH RNA detected in the same fraction. In all graphs shown, data are expressed as fold change relative to the levels of wild-type (WT) transcripts. The histograms depict mean and standard deviation of three independent experiments. The asterisk denotes statistically significant differences (Student's *t*-test, $*p<0.05$, $**p<0.01$).

As a third model we analysed cells from a patient with Xeroderma pigmentosum, an autossomic recessive condition characterized by increased sensitivity to ultraviolet irradiation and increased risk of skin cancer. It is caused by mutations in the *XPC* gene (MIM: 613208), which encodes a protein required for DNA repair (Khan, Oh et al. 2006, Khan, Oh et al. 2009). The cell line analysed is homozygous for two distinct mutations at the 3' splice site of intron 11 (**Table 1 and Figure 4A**). These mutations lead to skipping of exon 12, retention of intron 11 and activation of a 3' cryptic splice site in exon 12, resulting in introduction of PTCs (Khan, Oh et al. 2006). Quantitative real-time RT-PCR using the primers indicated in **Figure 4A** reveals a significant reduction in the total cellular levels of mutant *XPC* RNA compared to wild-type (**Figure 4B**). Analysis of sub-cellular fractions further shows that mutant transcripts are significantly less abundant in the cytoplasm, nucleoplasm and chromatin (**Figure 4C, D, E**).

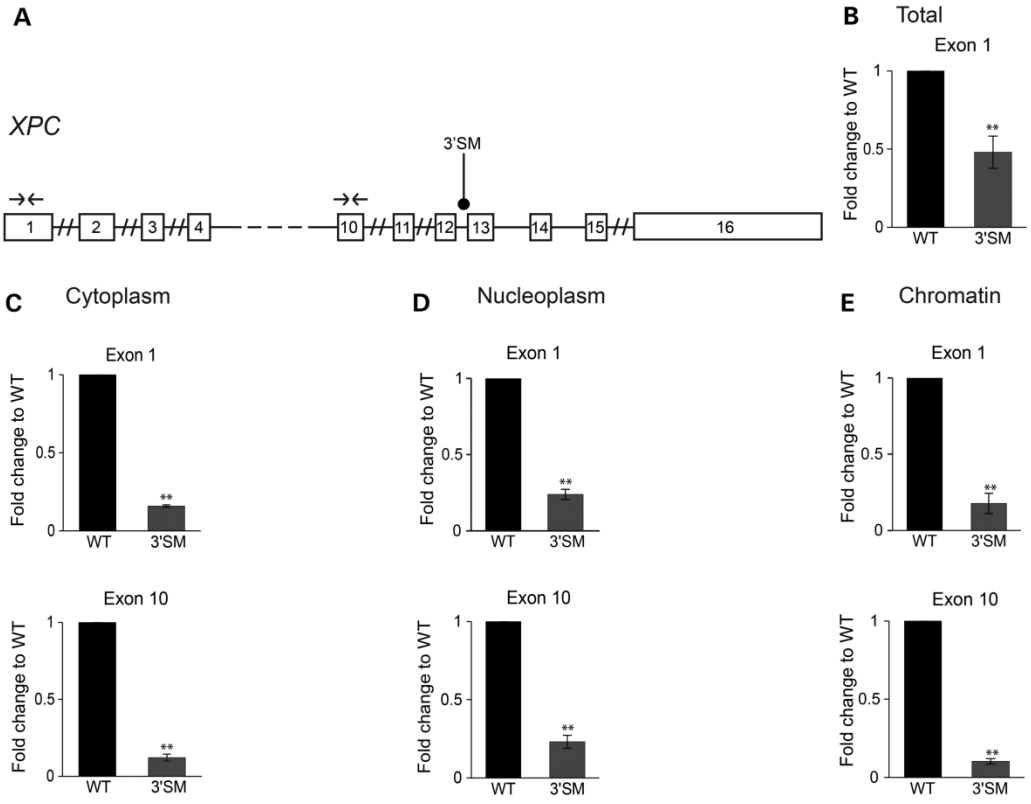


Figure 4. Sub-cellular distribution of wild-type and mutant *XPC* transcripts.
A) Illustration of the *XPC* gene structure (total length: 33525 bp). Exons are represented by numbered boxes and introns by lines; doubled intersected lines denote introns with more than 1000 bp. Gene region from exon 5 to exon 9 is represented by

a dashed line. Positioning of the mutation (3'SM) and primers used for PCR amplification (paired arrows) are indicated. B) Total cellular RNA was extracted from the indicated cell lines, reverse transcribed with random primers and analysed by RT-qPCR using primers for exon 1. The amount of PCR product obtained from each cell line was normalized to the level of GAPDH RNA detected in the same line. C) RNA was extracted from sub-cellular fractions isolated from each cell line and analysed by RT-qPCR amplified using primer sets for exon 1 and exon 10. The amount of PCR product obtained from each fraction was normalized to the level of GAPDH RNA detected in the same fraction. In all graphs shown, data are expressed as fold change relative to the levels of wild-type (WT) transcripts. The histograms depict mean and standard deviation of three independent experiments. The asterisk denotes statistically significant differences (Student's *t*-test, $**p<0.01$).

Altogether these results show that splicing mutations are consistently associated with reduced mRNA levels in the cytoplasm and, for a subset of mutations, down-regulation of expression is also detected in the nucleus. In contrast, mRNAs resulting from point mutations that introduce a PTC but do not interfere with splicing appear exclusively down-regulated in the cytoplasm.

3.1.3. A subset of genes carrying splicing mutations are less efficiently transcribed

To determine whether lower steady-state RNA levels in the nucleus result from reduced transcription of genes containing splicing mutations, we measured newly transcribed RNA levels by metabolic labelling with the natural uridine derivative 4-thiouridine (4sU). This approach provides direct access to newly synthesized transcripts with minimal toxic effects (Windhager, Bonfert et al. 2012), although it may induce a nucleolar stress response (Burger, Muhl et al. 2013). Nascent RNA was labelled by adding 4sU to the cell culture medium for 10 minutes followed by isolation of total cellular RNA. Newly transcribed RNA species containing thiol-groups were then biotinylated, purified using streptavidin-coated beads, and analysed by RT-qPCR (**Figure 5A**). As RNAPII transcribes with elongation rates ranging between 0.5 and 4 kb/min (Jonkers, Kwak et al. 2014), synthesis of new *TAZ* RNAs may take from 2.5 to 20 minutes, whereas *MARVELD2* and *XPC* RNAs may require between 7 or 8 minutes to approximately 1 hour. Thus, we expect that after incubation with

4sU for 10 minutes, most labelled RNAs are in the process of being synthesized and therefore should be confined to the chromatin fraction. The results shown in **Figure 5B** are in very good agreement with this prediction. To assess the extent to which transcription of the *TAZ*, *MARVELD2* and *XPC* genes differs between lymphoblastoid cell lines derived from normal individuals, we analysed a recently reported microarray dataset (Duan, Shi et al. 2013). The results show that the transcription rate of these genes is similar across cells from three distinct individuals (**Figure 5C**). Next, we compared the levels of nascent transcripts produced by wild-type and mutant genes using primers to amplify both exonic and intronic regions of *TAZ* (**Figure 5D**), *MARVELD2* (**Figure 5E**) and *XPC* (**Figure 5F**) transcripts. A significant down-regulation of nascent transcripts is observed for the *TAZ* 5' splice site (**Figure 5D, 5'SM**) and *MARVELD2* 3' splice site (**Figure 5E, 3'SM**) mutants, strongly suggesting that these genes are less efficiently transcribed. No evidence for reduced transcriptional activity of the *XPC* 3' splice site mutant gene is observed, arguing that the lower steady state RNA levels detected in the chromatin fraction likely reflect rapid nuclear degradation of these transcripts.

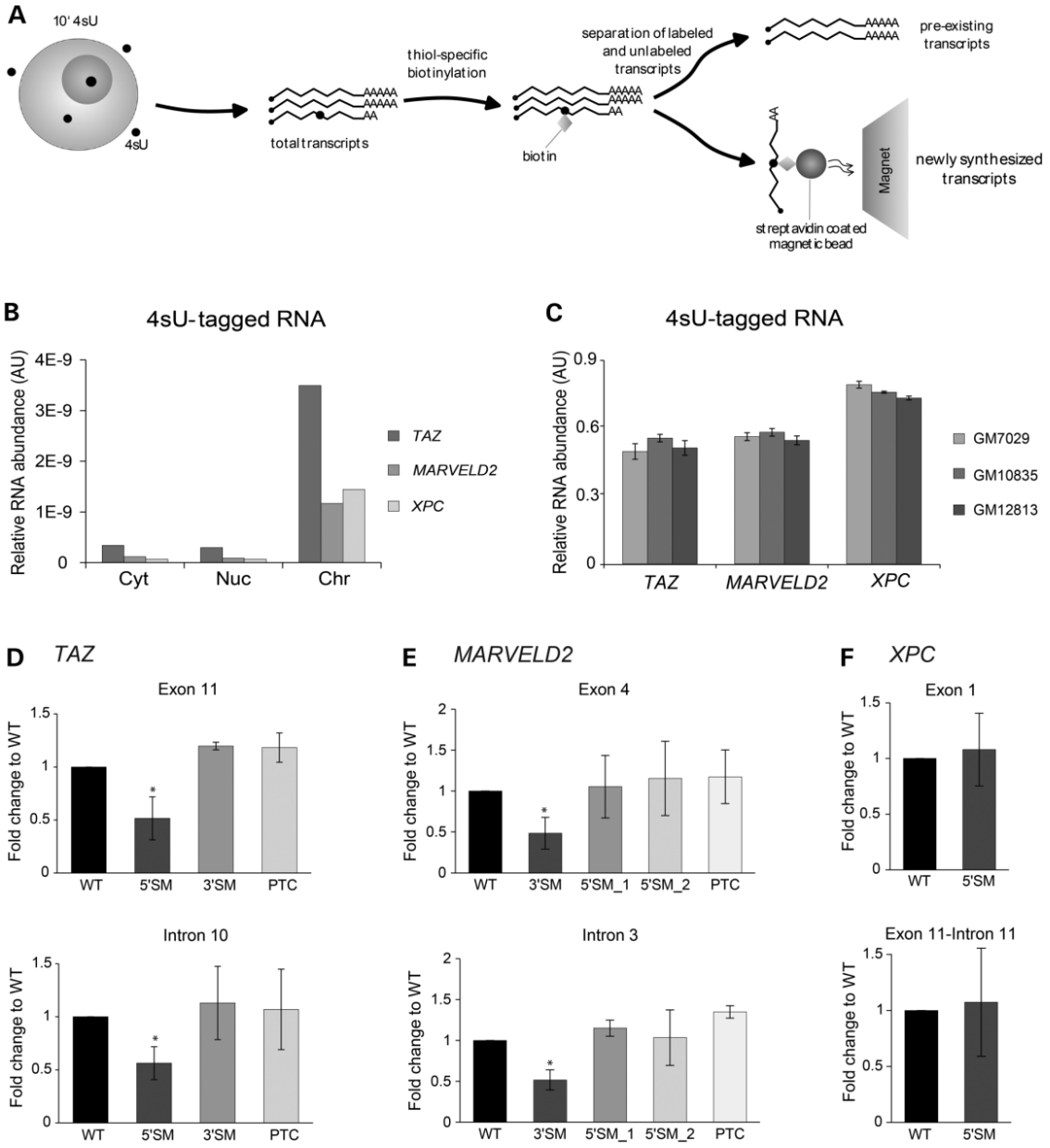


Figure 5. Analysis of nascent RNA by metabolic labelling.

A) Illustration of the metabolic labelling procedure. Cells in culture are incubated with 4-thiouridine (4sU). Total cellular RNA is extracted and thiol-containing molecules are biotinylated. Biotinylated RNA is then purified using streptavidin coated magnetic beads. B) Sub-cellular localization of 4sU-tagged RNA. Cells from a healthy donor (WT) were incubated with 4sU for 10 minutes and fractionated (Cyt: cytoplasm; Nuc: nucleoplasm; Chr: chromatin). RNA tagged with 4sU was purified from each fraction and analysed by RT-qPCR as described in figures 2, 3 and 4. AU (arbitrary units). C) Inter-individual differences of 4sU-tagged RNA. Nascent RNAs were isolated from lymphoblastoid cell lines derived from three unrelated healthy individuals (GM7029, GM10835 and GM12813) after incubation with 4sU for 2 hours (analysis of GSE34204 dataset, (Duan, Shi et al. 2013)). The amount of labelled *TAZ*, *MARVELD2* and *XPC* RNA was normalized to the level of labelled *GAPDH* RNA detected in the same cell line. The histogram depicts mean and standard deviation of three biological replicates

(independent cell cultures). AU (arbitrary units). D-F) Quantification of nascent transcripts produced by wild-type and mutant genes. Cells were incubated with 4sU for 10 minutes. Total 4sU-tagged RNA was purified and analysed by RT-qPCR using primers that recognize exonic (top) or intronic (bottom) regions. The amount of PCR product in each cell type was normalized to the level of *GAPDH* RNA detected in the same cell type. Data are expressed as fold change relative to the levels of wild-type (WT) transcripts. The histograms depict mean and standard deviation of three independent experiments. The asterisk denotes statistically significant differences (Student's *t*-test, $*p<0.05$).

3.1.4. NMD does not contribute to the observed down-regulation of mutant RNAs in the nucleus of patient-derived cell lines

To determine the contribution of NMD to the observed down-regulation of mutant RNAs in each sub-cellular fraction, cells were treated with cycloheximide (CHX), a drug that inhibits translation and hence indirectly blocks NMD (Schneider-Poetsch, Ju et al. 2010). After 3 hours of treatment, cells were fractionated and changes in RNA levels analysed by RT-qPCR. RNA levels in each treated fraction (CHX+) are expressed as fold change relative to the levels in the corresponding non-treated fraction (CHX-; **Figure 6, 7 and 8, A**). Alternatively, mutant RNA levels in each treated fraction (CHX+) are expressed as fold change relative to the levels of wild-type transcripts in the corresponding fraction from treated cells (**Figure 6, 7 and 8, B**). Analysis of TAZ (**Figure 6**), MARVELD2 (**Figure 7**) and XPC (**Figure 8**) mutant and wild-type transcripts shows that treatment with CHX consistently results in an increase in RNA levels in the cytoplasm. This increase is most obvious for mutant transcripts, as expected since their degradation by NMD is most probably impaired by CHX. An exception is the MARVELD2 PTC mutant, which gives rise to RNAs that are not affected by CHX, suggesting that they escape NMD. Accordingly, this particular mutant has been described to encode a truncated tricellulin protein (Riazuddin, Ahmed et al. 2006). The finding that CHX induces accumulation of wild-type transcripts is also in agreement with previous reports (Rajavel and Neufeld 2001) (Lareau, Inada et al. 2007).

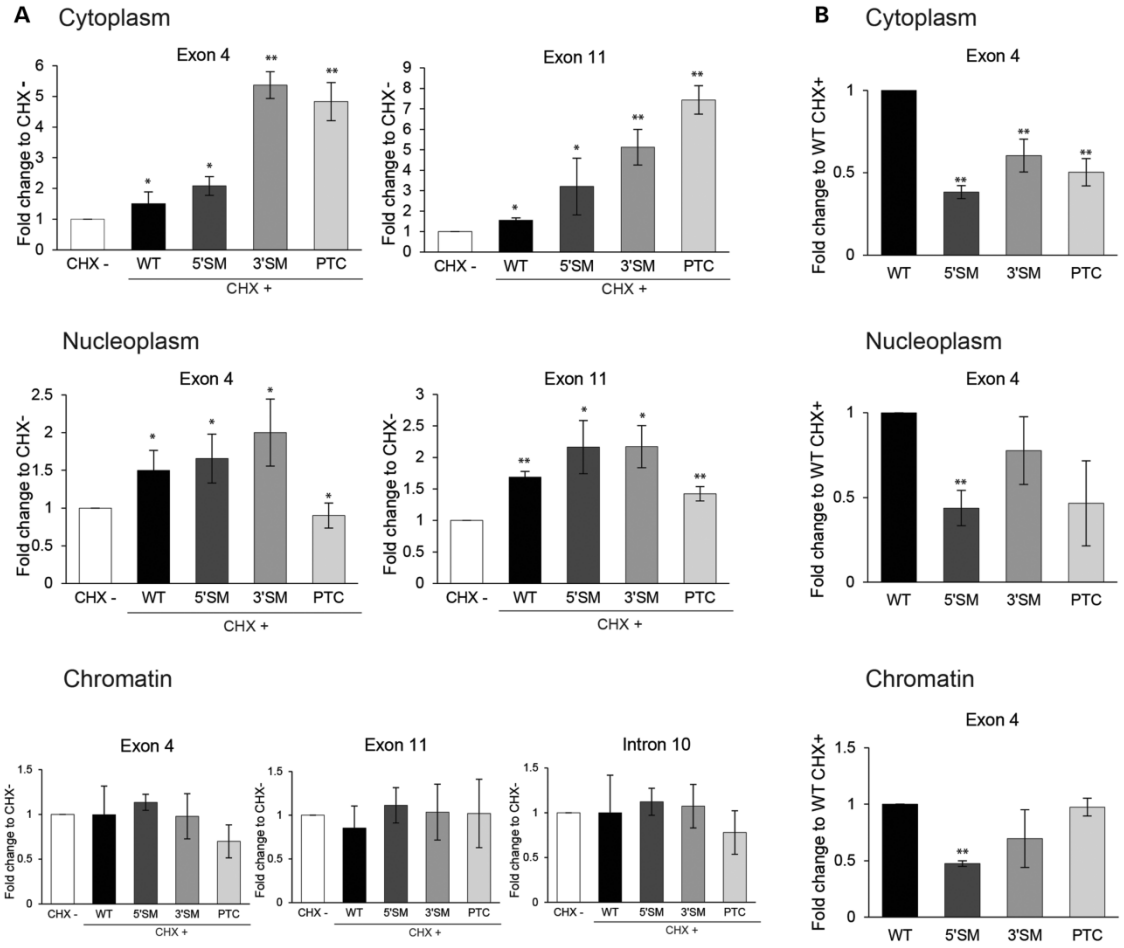


Figure 6. Effect of cycloheximide on *TAZ* transcripts.

Cells were either non-treated (CHX-) or treated with cycloheximide for 3 hours (CHX+). The levels of WT and mutant transcripts in each sub-cellular fraction were analysed by RT-qPCR using the indicated primer sets. The amount of PCR product was always normalized to the level of *GAPDH* RNA. Data are expressed as fold change relative to the levels of non-treated cells (A) or as fold change relative to the levels of wild-type transcripts in treated cells (B). The histograms depict mean and standard deviation of three independent experiments. The asterisk denotes statistically significant differences (Student's *t*-test, $*p<0.05$, $**p<0.01$).

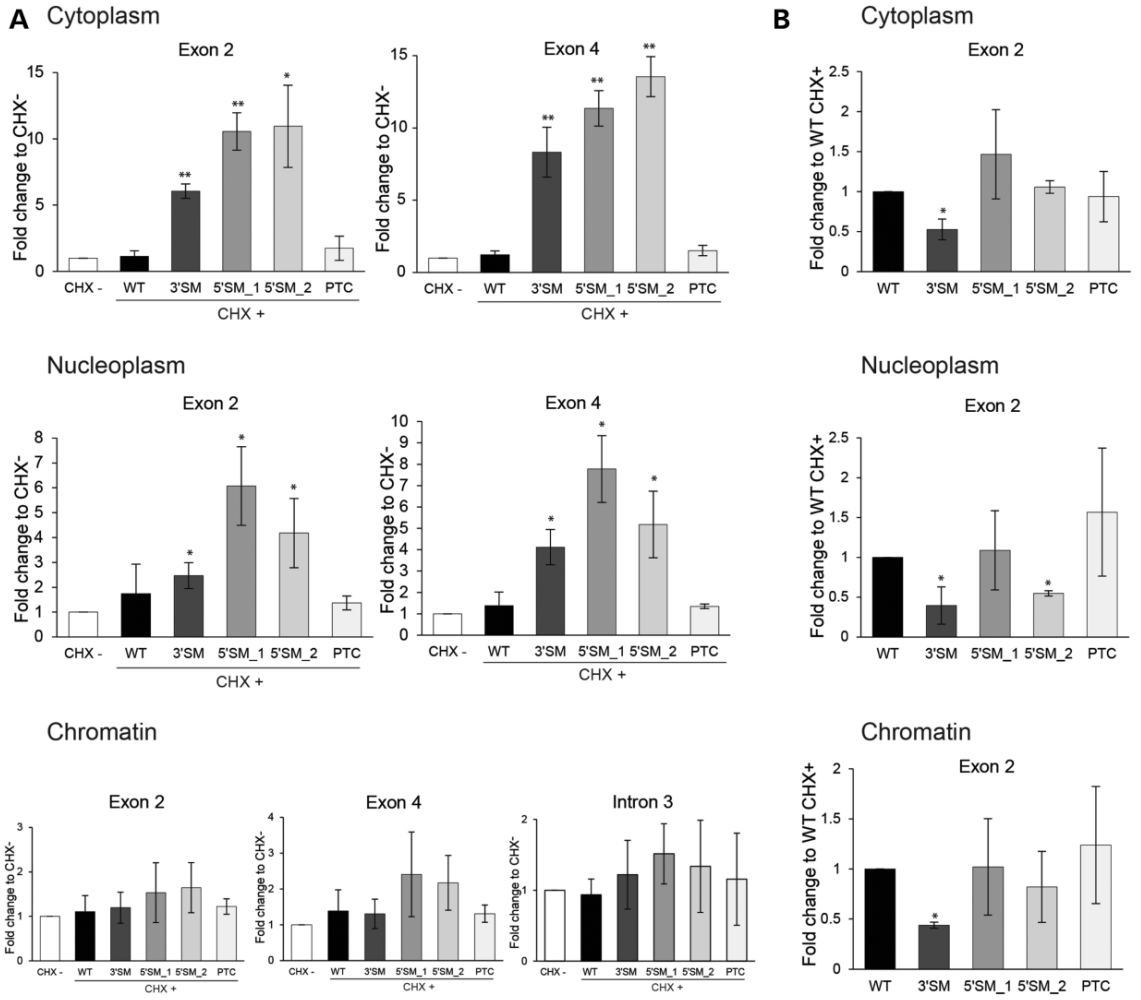


Figure 7. Effect of cycloheximide on *MARVELD2* transcripts.

Cells were either non-treated (CHX-) or treated with cycloheximide for 3 hours (CHX+). The levels of WT and mutant transcripts in each sub-cellular fraction were analysed by RT-qPCR using the indicated primer sets. The amount of PCR product was always normalized to the level of *GAPDH* RNA. Data are expressed as fold change relative to the levels of non-treated cells (A) or as fold change relative to the levels of wild-type transcripts in treated cells (B). The histograms depict mean and standard deviation of three independent experiments. The asterisk denotes statistically significant differences (Student's *t*-test, $*p<0.05$, $**p<0.01$).

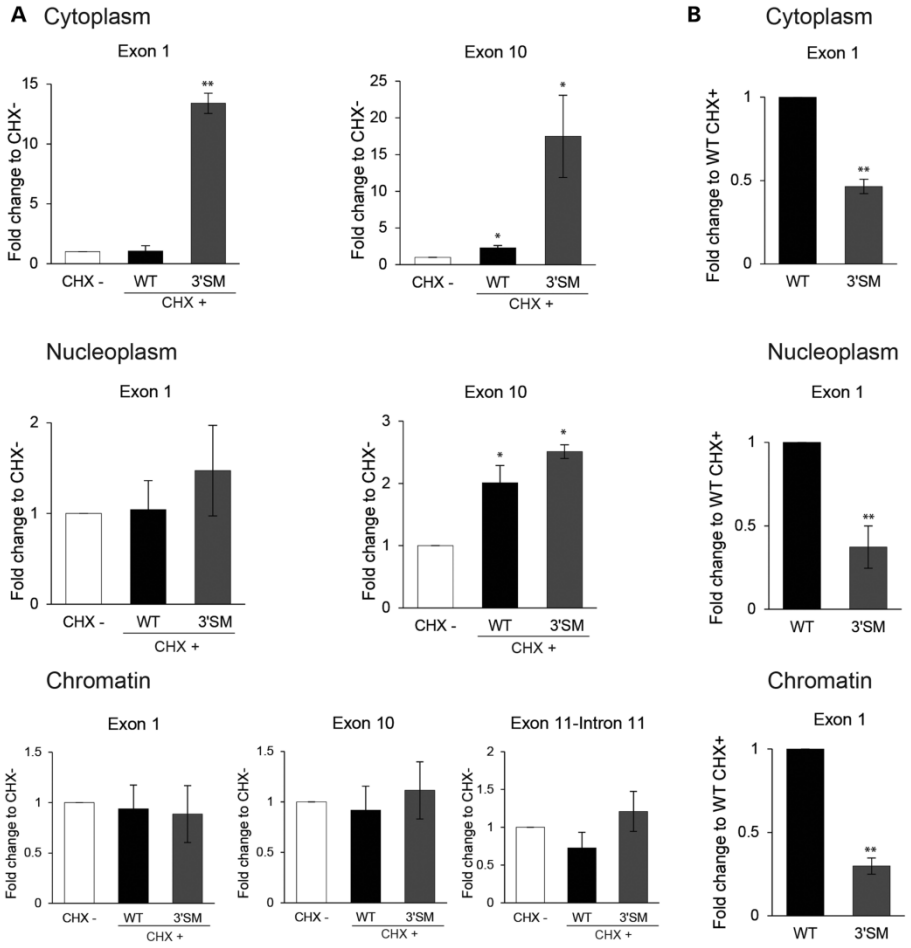


Figure 8. Effect of cycloheximide on *XPC* transcripts.

Cells were either non-treated (CHX-) or treated with cycloheximide for 3 hours (CHX+). The levels of WT and mutant transcripts in each sub-cellular fraction were analysed by RT-qPCR using the indicated primer sets. The amount of PCR product was always normalized to the level of *GAPDH* RNA. Data are expressed as fold change relative to the levels of non-treated cells (A) or as fold change relative to the levels of wild-type transcripts in treated cells (B). The histograms depict mean and standard deviation of three independent experiments. The asterisk denotes statistically significant differences (Student's *t*-test, $*p<0.05$, $**p<0.01$).

An accumulation of both wild-type and mutant RNAs is further detected in the nucleoplasm of CHX treated cells. This observation argues that the lower steady state levels of mutant transcripts observed in association with the nucleoplasm without a corresponding decrease in the chromatin fraction could be due to contamination of the nucleoplasmic fraction by mRNAs that have already been exported from the nucleus but remain associated with the cytoplasmic side of the nuclear envelope, as previously proposed (Popp and

Maquat 2013). In contrast, CHX does not significantly alter the levels of wild-type and mutant RNAs associated with the chromatin fraction. However, the levels of *TAZ* 5'SM, *MARVELD2* 3'SM and *XPC* 3'SM RNAs persist reduced compared to wild-type in the chromatin fraction of CHX treated cells (**Figure 6, 7 and 8, B**). Noteworthy, *TAZ* 5'SM and *MARVELD2* 3'SM RNAs, which are less efficiently transcribed (**Figure 5D, E**), respond less to CHX treatment than other mutant forms of the same gene. The mild effect of CHX on cytoplasmic levels of *TAZ* 5'SM transcripts is in agreement with the finding that the majority of these RNAs are devoid of PTCs and therefore should not be degraded by NMD. Taken together, these observations suggest that some splicing mutations result in RNAs that are primarily degraded by NMD in the cytoplasm, while others can be targeted by transcription-coupled quality control mechanisms that operate independently from NMD.

3.2. Deep-intronic mutations and human disease

Part of the data presented in this section is published in the *Human Genetics* peer-reviewed journal. The article in the publication format can be found in the Appendix of this thesis.

Rita Vaz-Drago, Noélia Custódio and Maria Carmo-Fonseca

Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa,

1649-028 Lisboa, Portugal

Author contribution

Rita Vaz-Drago, Noélia Custódio and Maria Carmo-Fonseca wrote the manuscript. Rita Vaz-Drago reviewed most of the literature and performed data analysis. All authors revised the entire manuscript.

3.2.1. Overview

Having focused before on mutations that cause disease by disrupting canonical splice sites, here we describe the impact of mutations that create non-canonical splice sites deep within introns.

Although next-generation sequencing has revolutionized clinical diagnostic testing, sequence information restricted to exons and exon-intron boundaries fails to identify the genetic cause of the disease for a substantial proportion of patients. Current estimates indicate that only 1-5% of all disease-causing mutations are expected to fall more than 100 base pairs away from exon-intron junctions. Yet, non-coding intronic regions account for 95% of total protein-coding gene length and are increasingly being described as important players in gene expression regulation. In this context we decided to reviewed evidence from mRNA analysis and entire genomic sequencing indicating that pathogenic mutations can occur deep within the introns of over 75 disease-associated genes. DNA variants located deep within introns most commonly lead to pseudo-exon inclusion due to creation of non-canonical splice sites followed by the activation of a pre-existing non-canonical splice sites or changes in splicing regulatory elements. Additionally, deep intronic mutations can disrupt transcription regulatory motifs and non-coding RNA genes. This chapter aims to highlight the importance of studying variation in deep intronic sequence as a cause of monogenic disorders as well as hereditary cancer syndromes.

3.2.2. Human introns are 20 times longer than exons

Introns represent approximately 25% of the human genome and are part of 95,5% of all human genes (Louhichi, Fourati et al. 2011). Size distribution of human exons and introns showed that almost 50% of introns are bigger than 2000 bp, as opposed to exons that are significantly shorter, ranging between 100-200 bp (**Figure 9**).

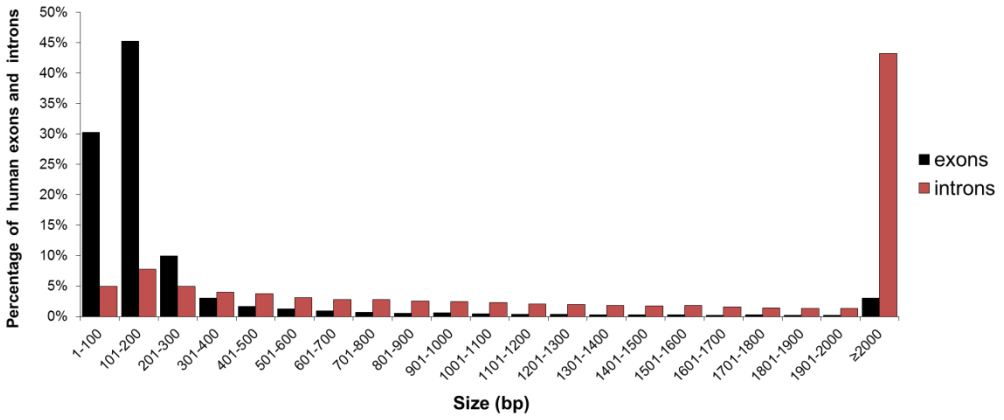


Figure 9. Size distribution of human exons and introns.
Analysis was performed using 58529 annoted protein coding transcripts (GRCh38). A total of 541182 exons and 490055 introns were size-distributed in 100 bp intervals.

3.2.3. Deep intronic mutations most often lead to the creation of novel, non-canonical donor splice sites

Because introns are removed from nascent transcripts during pre-mRNA processing, intronic sequences in genes have been considered as "junk DNA". However, the description of many disease-associated genetic variants located within introns often far away from the splice-junctions (Xiong, Alipanahi et al. 2015, Hsiao, Bahn et al. 2016) constitute an evidence for intron functionality. To date, mutations in deep intronic regions have been documented in multiple diseases. We reviewed 117 studies published between 1983 and 2016 describing 185 intronic mutations located at least 100 bp from the nearest canonical splice site, across 77 different disease genes (**Figure 10**).

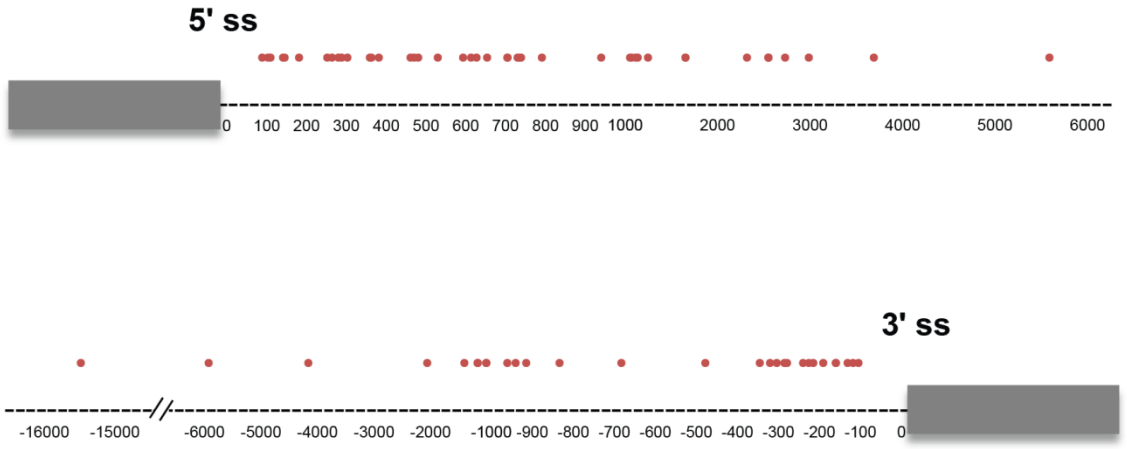


Figure 10. Distribution of deep-intronic mutations across introns.
The location of all deep-intronic mutations reported since 1983 and reviewed in this study are represented relative to canonical 5' and 3'splice sites (5' ss, 3' ss).

The most common consequence of this type of mutation involves the creation of a non-canonical donor splice site and subsequent activation of a pre-existing acceptor non-canonical splice site (**Figure 11** and **Figure 12A**). Less frequently a deep-intronic mutation creates a novel acceptor splice site and activates a downstream non-canonical donor splice site (**Figure 11** and **Figure 12B**). This combined creation and activation of non-canonical splice sites triggers the inclusion of a pseudo-exon in the mutant mRNA. Disease-associated pseudo- or cryptic exons range in size from 30 to 344 base pairs (**Figure 13**). The appearance of a pseudo-exon generally disrupts the reading frame introducing a premature termination codon that targets the mutant mRNA for degradation by nonsense mediated decay (NMD) (Popp and Maquat 2013).

Most deep intronic mutations have no effect on canonical splice sites. Yet, some mutations that create a new splice site interfere with recognition of natural splice sites. Weakening of canonical splice sites is frequently observed when deep intronic mutations are less than 150 bp away from the natural exon-intron junctions.

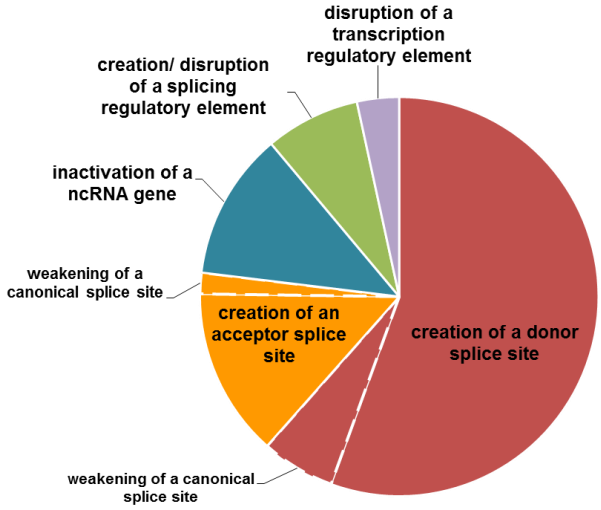


Figure 11. Proportion of mechanisms by which deep-intronic mutations alter gene expression.

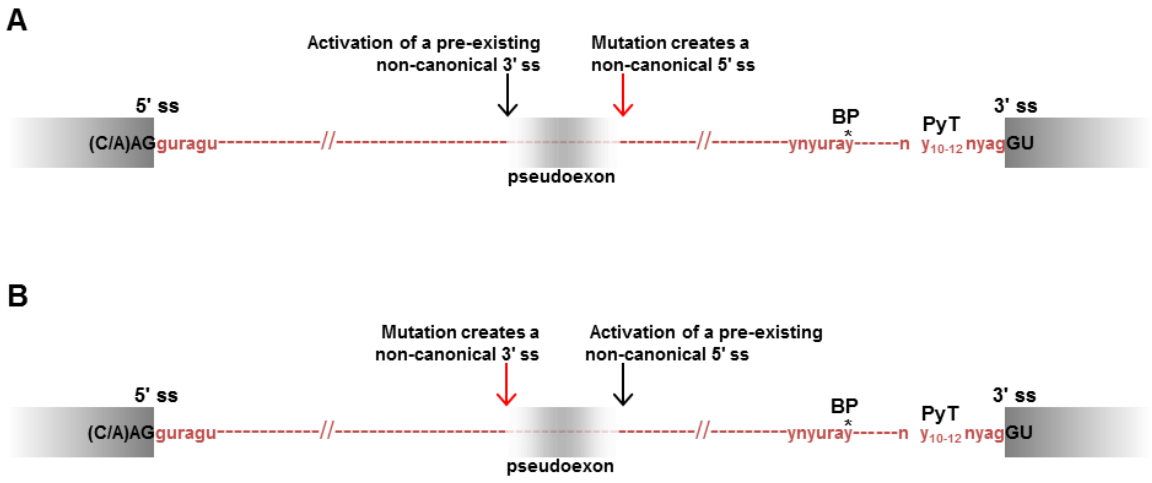


Figure 12. Pseudo-exon inclusion triggered by mutations that create non-canonical splice sites.

A) Illustration of a mutation that creates a novel donor (5' ss) splice site and activates a pre-existing acceptor (3' ss). B) Illustration of a mutation that creates a novel acceptor (3' ss) splice site and activates a pre-existing donor (5' ss).

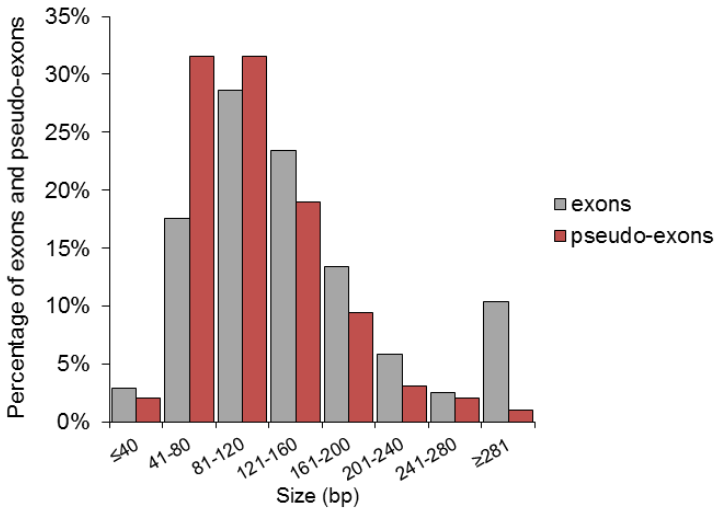


Figure 13. Size distribution of pseudo-exons.
Pseudoexons referred to in Tables 1, 2 and 3 were size-distributed in 40 bp intervals. For comparison, the size distribution of authentic middle exons (excluding first and last exons) is indicated. Size and frequency of authentic human exons (*h19*) were calculated using BED files downloaded from the UCSC *Table Browser*.

Inactivation of ncRNA genes is the second most representative class and account for 12% of all the reviewed deep-intronic mutations (**Figure 11**). Among those, point mutations in the *RNU4ATAC* gene account for the majority of the cases. The *RNU4ATAC* gene, which codes for the minor spliceosomal U4atac snRNA, is located within intron 2 of the protein-coding *CLASP1* gene, 682 to 556 bp upstream of exon 3 (Edery, Marcaillou et al. 2011). Consistent with loss-of-function of the mutant snRNA, higher levels of unspliced U12-type introns were detected in patient-derived fibroblasts.

Alternatively, 10% of disease-causing deep intronic mutations alter the binding of RNA- or DNA-binding proteins (**Figure 11**). In the first case, these type of mutation can either create or disrupt splicing enhancer or silencer elements, respectively, priming the inclusion of a pseudo-exon in the mutant mRNA. Less frequently, deep intronic mutations can deregulate transcription of the mutant gene by disrupting transcription regulatory motifs, most often placed within first introns.

3.3. RNA metabolic labelling introduces bias in splicing analysis

Rita Vaz-Drago, Noélia Custódio, Célia Carvalho and Maria Carmo-Fonseca

Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, 1649-028 Lisboa, Portugal

Author contribution

Rita Vaz-Drago, Noélia Custódio and Maria Carmo-Fonseca designed the experiments. Rita Vaz-Drago performed all the experiments and data analysis. Célia Carvalho and Noélia Custódio designed most of the primers and contributed with expertise.

3.3.1. Overview

An important step toward understanding gene regulation in health and disease is the elucidation of the kinetics of splicing, since disruption of splicing is one of the major causes of monogenic disorders.

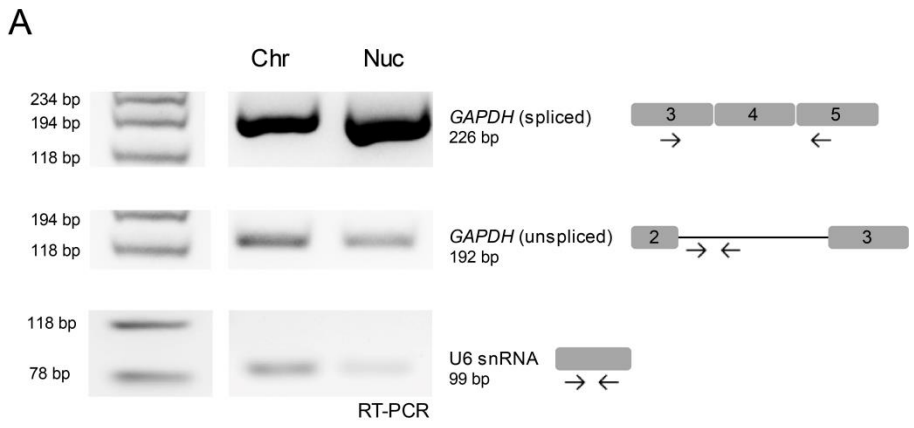
A variety of approaches have been used to determine efficiency of splicing. Evidence from many studies suggested that most splicing occur co-transcriptionally, thus purification of newly synthesized transcripts can greatly increases the accuracy of the determination of splicing efficiency. Recent studies used RNA metabolic labelling with short pulses of 4-thiouridine (4sU) or 4-thiouracil (4tU) to isolate newly transcribed molecules and determine kinetics of pre-mRNA splicing. This approach relies on treatment with a thio-reactive reagent to biotinylate the tagged RNA, which is then affinity-purified with streptavidin. A limitation of this method is that the reaction of 4sU with the commonly used biotin-HPDP is inefficient, which may lead to an over-representation of longer RNA molecules in the purified fraction. In this chapter, we show that the 4sU labelling does not interfere with splicing efficiency. However, nascent RNA purified with biotin-HPDP contains a significantly higher proportion of unspliced long introns compared to RNAs purified with the more efficient biotinylation strategy that uses methanethiosulfonate (MTS) reagent. Thus, the splicing kinetics of long introns may be selectively under-estimated in studies using biotin-HPDP.

3.3.2. 4sU incorporation does not interfere with splicing efficiency

Previous studies have shown that incorporation of 4sU into RNA causes minimal interference to cell growth and gene expression (Melvin, Milne et al. 1978, Cleary, Meiering et al. 2005, Kenzelmann, Maertens et al. 2007, Dolken, Ruzsics et al. 2008, Friedel and Dolken 2009, Amorim, Cotobal et al. 2010). Yet, incubation with 4sU for 48 hours at concentrations ranging between 50 μM and 500 μM may affect cell viability (Tani and Akimitsu 2012) and incubation with 100 μM 4sU for 6 hours causes nucleolar stress and inhibits rRNA synthesis (Burger, Muhl et al. 2013). Whether incorporation of 4sU into nascent transcripts interferes with RNA processing remains unclear. To investigate a potential effect of 4sU-tagging on human pre-mRNA splicing, HEK 293 cells were grown in the presence of 500 μM 4sU for 2 and 60 minutes; splicing efficiency of selected transcripts was then measured in the chromatin and nucleoplasm fractions. It is well established that chromatin-associated RNA is enriched in nascent transcripts still attached to the RNAPII, whereas RNA from the nucleoplasm represents predominantly transcripts that have already been released from the DNA template and are in transit to the cytoplasm (Wuarin and Schibler 1994, Dye, Gromak et al. 2006, Pandya-Jones and Black 2009). To assess our fractionation procedure, chromatin-associated and nucleoplasmic-released transcripts were reverse transcribed with random primers and PCR amplified using primers for spliced and unspliced *GAPDH* RNA and U6 RNA (**Figure 14A**). To detect *GAPDH* mRNA, a forward primer was designed to bind exon 3 and a reverse primer was designed to bind exon 5, ensuring that only the spliced isoform was amplified. To specifically amplify *GAPDH* pre-mRNA a forward and a reverse primer were designed to bind intron 2. After completing the RT-PCR, equal amounts of amplicon from each of the fractionated samples were separated in a 1% agarose gel. As expected, GAPDH mRNA is detected in both chromatin and nucleoplasm fractions whereas pre-mRNA is detected mainly in the chromatin fraction (**Figure 14A**). The U6 spliceosomal snRNA is predominantly detected in the chromatin fraction (**Figure 14A**), as previously reported (Tilgner, Knowles et al. 2012).

For splicing efficiency analysis, we selected genes, based on three criteria: level of expression, intron length and intron class. First, genes that were highly expressed in HEK 293 cells were selected based on publicly available RNA-seq

data. Second, we analysed both long (13-89 kbp) and short (240 bp) introns. Third, we analysed a U12-dependent intron in the *CTNNBL1* gene that was previously shown to be slowly spliced (Singh and Padgett 2009). Unspliced transcripts were detected using PCR primers flanking exon-intron junctions, whereas spliced transcripts were detected with primers located in either neighboring exons or spanning the junction of three exons, depending on the length of the introns. To evaluate the linearity of each primer pair, we performed a 10-fold or 5-fold dilution series of template cDNA (**Figure 14B**). Threshold cycles obtained in the RT-qPCR reactions were plotted against each of the cDNA dilutions and the slope of the trendline was determined by linear regression. Primer efficiency was then calculated using the equation $E = (10^{(-1/slope)}-1) \times 100$, where E is the efficiency. All primers presented efficiencies ranged from 90% to 110%. The difference between efficiencies of U6 primers (internal control) and each one of the spliced and unspliced isoforms of *GAPDH*, *HPRT1*, *COL4A6* and *CTNNBL1*, as well as the difference between efficiencies of spliced and unspliced transcripts are never higher than 10%, as recommended (Schmittgen and Livak 2008). Specificity of the primers was confirmed by melting curve analysis. For all pairs studied we could observe a single sharp peak, suggesting that a specific product was amplified during the RT-qPCR reactions (**Figure 14B**).



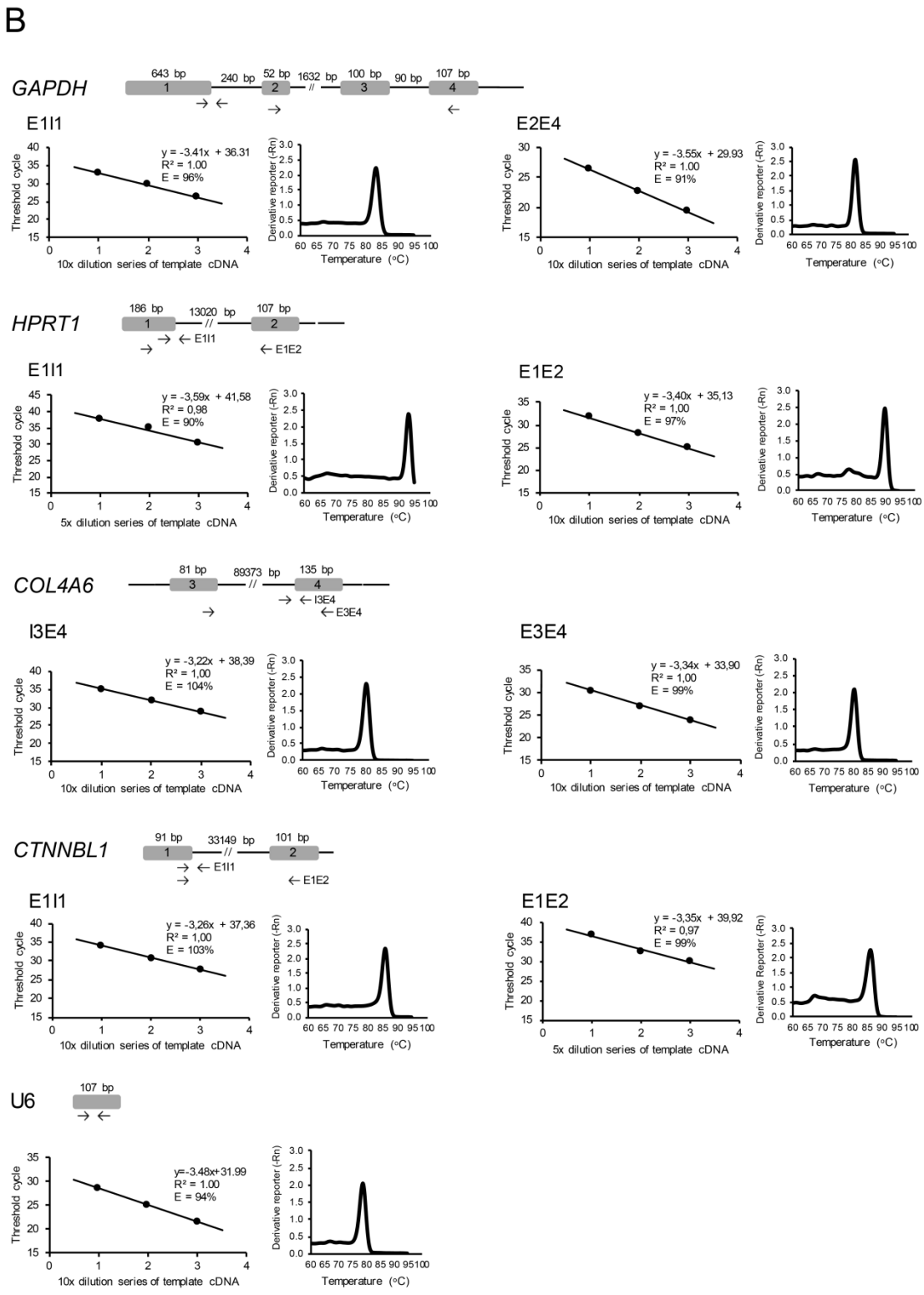


Figure 14. Efficiency of cellular fractionation and RT-qPCR reactions.

A) Efficiency of the fractionation protocol was evaluated by performing RT-PCR analysis. RNA was isolated from HEK 293 cells, reverse transcribed with random primers and PCR amplified using primers for spliced and unspliced *GAPDH* RNA and total U6 RNA. Equal amounts of cDNA from total and fractionated samples were loaded

per lane. DNA size markers in base pairs (bp) are indicated on the left. Expected fragment sizes (bp) are indicated in parentheses base pairs. B) Linearity and specificity of primers were evaluated by RT-qPCR. Illustration of all the studied genes is represented in association with each plot. Exons are represented by numbered boxes and introns by line. Primers used for PCR amplification (paired arrows) of unspliced and spliced isoforms are also shown.

To determine whether the incorporation of 4sU into nascent transcripts affects pre-mRNA splicing, equal amounts of RNA were taken from the chromatin and nucleoplasm fractions, then reverse transcribed with random primers and PCR amplified using specific primer pairs for spliced and unspliced transcripts, as indicated in **Figure 14B**. The abundance of PCR product was normalized to the level of U6 snRNA detected in each fraction and in the same RT-qPCR run.

The isolation of chromatin-associated transcripts comprehends the use of detergent, salt and urea, to assure that only nascent, RNA-bound transcripts are released from the intact chromatin pellet after DNase treatment. The purification of these RNA isolates is readily completed by phenol/chloroform extraction, which allows us to perform an unbiased and accurate measurement of steady-state RNA splicing ratios. Any changes in splicing efficiency due to 4sU incorporation would be reflected in an increase (or decrease) of the proportion of spliced transcripts relatively to the total number of isoforms, here presented as a measure of intron removal levels (**Figure 15**). The proportion of spliced products in each fraction was calculated as the ratio between the amount of spliced product and total amount of spliced and unspliced transcripts (**Figure 15**).

The results show that the vast majority (>95%) of *COL4A6*, *HPRT1* and *GAPDH* transcripts in the chromatin fraction are spliced, indicating that the analysed introns are efficiently excised from pre-mRNAs shortly after transcription (**Figure 15A, B, C**). In contrast, only 20% of transcripts containing the U12-dependent *CTNNBL1* intron are spliced in the chromatin, suggesting rather inefficient splicing; even in the nucleoplasm the intron is still detected in ~30% of transcripts (**Figure 15D**). Most important, we observe similar proportion of spliced transcripts in both chromatin and nucleoplasm fractions of cells that were either non-treated (**Figure 15A-D, -4sU**) or incubated with 4sU for 2 and

60 minutes. This reveals that under the conditions used in this study 4sU tagging does not interfere with pre-mRNA splicing efficiency.

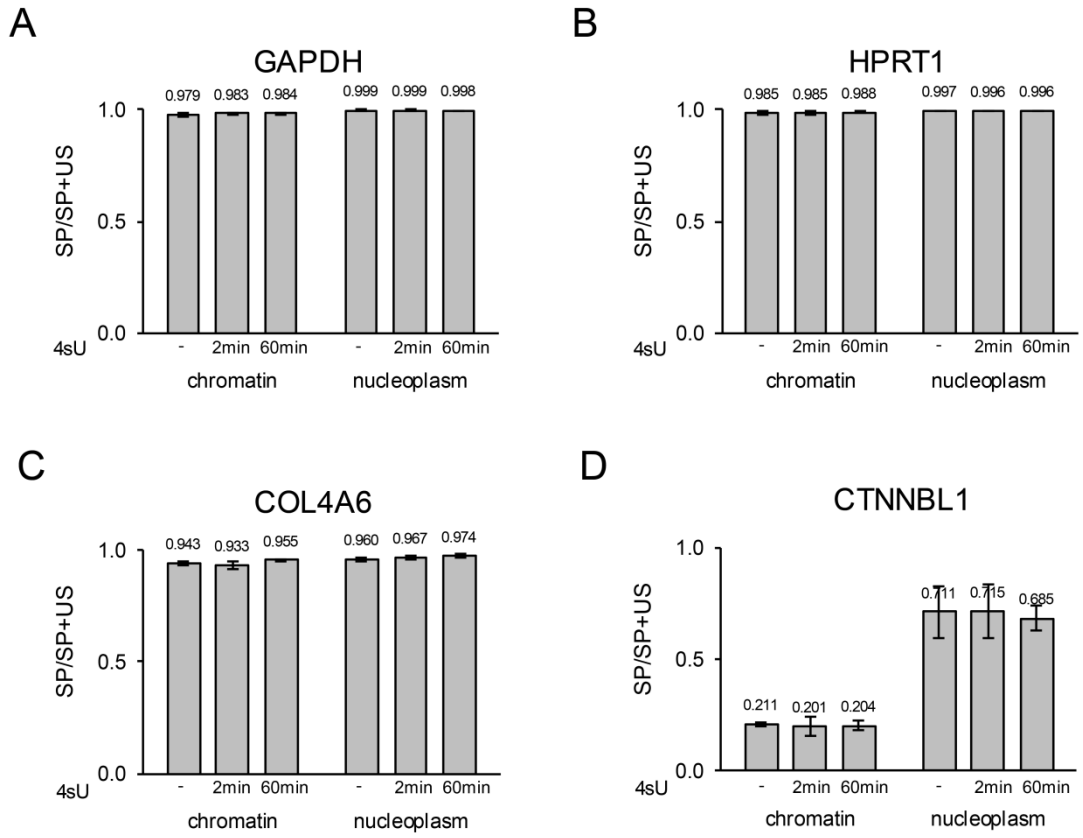


Figure 15. 4sU incorporation does not interfere with splicing.
Cells were either grown in the absence of 4sU (-) or incubated with 4sU for 2 and 60 minutes and RNA was extracted from chromatin and nucleoplasm fractions. For the indicated genes, the amount of spliced (SP) product detected by RT-qPCR and normalized to U6 RNA is shown as proportion relative to the total amount of spliced (SP) and unspliced (US) transcripts. Histograms depict mean and standard deviation of three independent experiments.

3.3.3. HPDP-biotin purification results in biased enrichment of long unspliced transcripts

To date, all metabolic RNA labelling studies addressing the kinetics of pre-mRNA splicing have used N-[6-(Biotinamido)hexyl]-3′-(2′-pyridyldithio)-propionamide (biotin-HPDP) to conjugate thiol groups incorporated in RNAs to biotin. However, the reaction and corresponding enrichment of 4sU-RNA with

HPDP are inefficient (Duffy, Rutenberg-Schoenberg et al. 2015). This results in a bias enrichment toward longer RNAs (Miller, Robinson et al. 2009, Miller, Schwalb et al. 2011) because smaller RNAs contain fewer uridine residues and therefore have lower probability of successful labelling. Such length bias may impact on splicing rate estimates, which are based on relative levels of transcripts of different sizes (splice and unspliced). Here, we investigated how length bias influences splicing dynamics estimates in human cells. We compared the proportion of transcripts containing spliced and unspliced introns of different length in 4sU-RNAs purified with biotin-HPDP and a more efficient biotinylation strategy that uses methanethiosulfonate (MTS) reagents (Duffy, Rutenberg-Schoenberg et al. 2015).

To determine whether MTS and HPDP chemistries influence splicing analysis, cells were first incubated with 4sU for 2 and 60 minutes and labelled transcripts were then separated from the pre-existing RNA using either HPDP-biotin (Dolken, Ruzsics et al. 2008) or MTS-biotin (Duffy, Rutenberg-Schoenberg et al. 2015). After the 2-minute pulse, significantly higher amount of RNA was recovered with the MTS-biotin purification approach compared to HPDP-biotin (**Figure 16**). In contrast, after the 60-minute pulse similar amount of RNA was obtained with both methods (**Figure 16**), suggesting that the lower efficiency of HPDP-biotin is compensated by the longer incubation pulse.

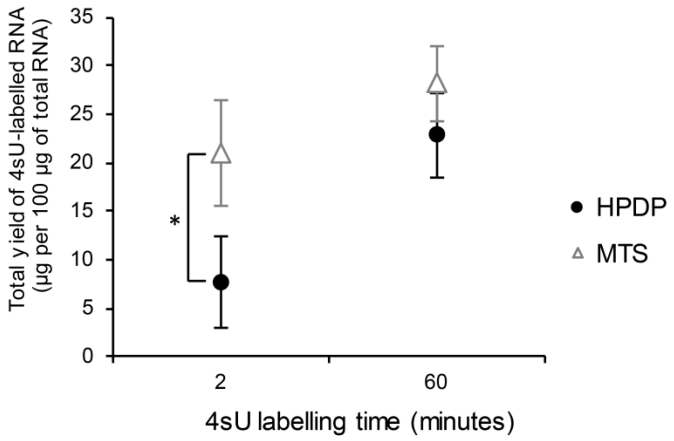


Figure 16. Comparation of RNA yields obtained with HPDP- and MTS-biotin.
Total yield of RNA recovered using HPDP- and MTS-biotin after incubation with 4sU for the indicated time. The graph depicts mean and standard deviation of three independent experiments. The asterisk denotes statistically significant differences (Student's *t*-test $*p<0.05$).

Next, we compared the proportion of splicing of 4sU-labeled transcripts purified with either HPDP-biotin or MTS-biotin. After the 2-minute pulse, purification of 4sU-labeled RNAs with both methods resulted in similar splicing proportions for the 240bp intron of *GAPDH* and the 13,020bp *HPRT1* intron (**Figure 17A**). However, significantly higher splicing values were estimated with MTS-biotin for the longer *COL4A6* and *CTNNBL1* introns (**Figure 17A**). After the 60-minute pulse, significantly higher splicing values were estimated with MTS-biotin for all introns analysed, although the difference is more striking for the longer *COL4A6* and *CTNNBL1* introns (**Figure 17B**). Thus, splicing efficiency is consistently under-estimated by HPDP-biotin.

Assuming that the purification of chromatin-associated transcripts does not introduce any length bias, we reasoned that by comparing the proportion of spliced transcripts between the chromatin-associated RNA fraction and both 4sU-labelled RNAs purification methods we could evaluate the dimension of the bias towards unspliced isoforms. Thus, we compared the proportion of splicing estimated in chromatin-associated transcripts and 4sU-labelled RNAs. After a 2-minute pulse the very long (89,373bp) *COL4A6* intron appears more efficiently spliced in the chromatin fraction (**Figure 17A**), suggesting that during such a short incubation with 4sU there is a bias enrichment of labelled RNAs toward unspliced molecules irrespective of using HPDP-biotin or MTS-biotin. Such bias is no longer observed after a pulse of 60 minutes, but only if labelled RNAs are purified with MTS-biotin (**Figure 17B**). A different scenario is observed for the U12-dependent *CTNNBL1* intron; in this case the proportion of spliced products is much lower in the chromatin fraction than in labelled RNAs purified with MTS-biotin, particularly after incubation with 4sU for 60 minutes (**Figure 17B**). This is most likely because during the 60-minute pulse many nascent transcripts are released from the chromatin and accumulate as spliced mRNAs in the cytoplasm.

Therefore, our work provides a valuable experimental test that shows that some methods of purification of nascent transcripts are susceptible to bias and highlights the importance of selecting the appropriate biochemical method accordingly with the type of analysis preformed.

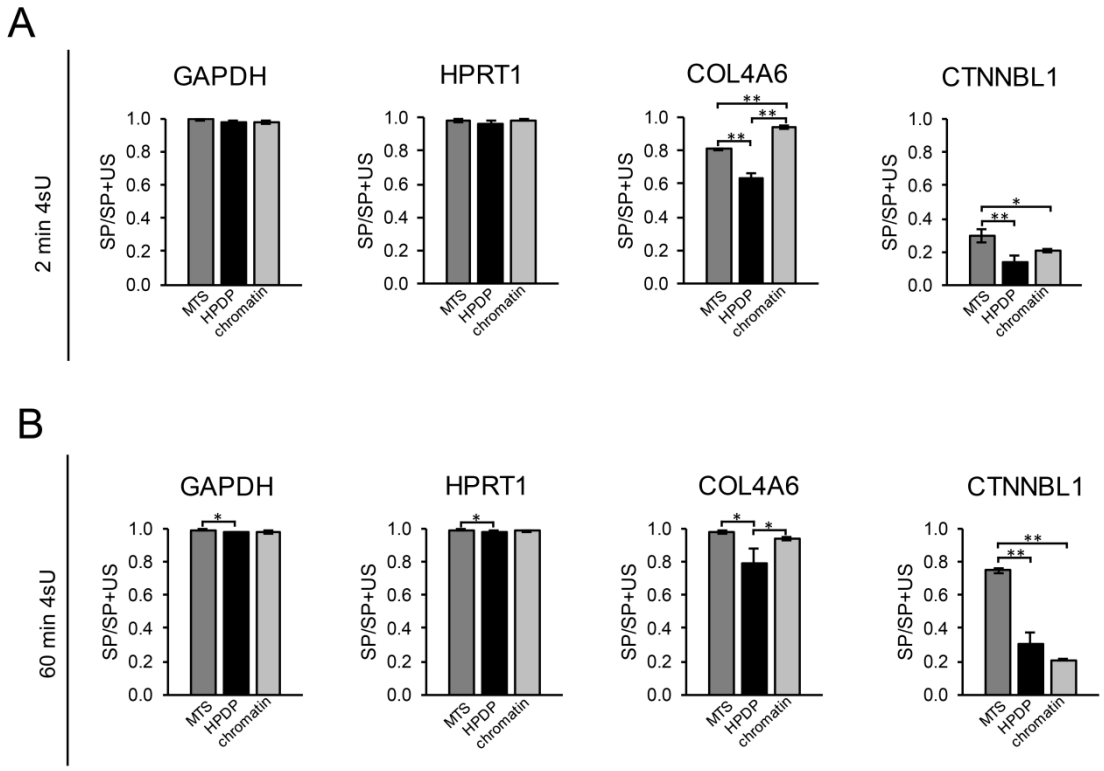


Figure 17. Biased enrichment of long unspliced transcripts purified with HPDP-biotin. 4sU-labelled RNA was purified with either HPDP- or MTS-biotin after pulses of A) 2 and B) 60 minutes. For comparison, chromatin-associated RNA was analysed from cells that were not incubated with 4sU. For the indicated genes, the amount of spliced (SP) product detected by RT-qPCR and normalized to U6 RNA is shown as proportion relative to the total amount of spliced (SP) and unspliced (US) transcripts. Histograms depict mean of three or four independent experiments. The asterisk denotes statistically significant differences (Student's *t*-test *$p<0.05$; **$p<0.01$).

3.4. Kinetics of pre-mRNA cleavage and termination in living cells

Rita Vaz-Drago, Ana C de Jesus[§], Robert M Martin, Célia Carvalho, José Rino, Noélia Custódio, Maria Carmo-Fonseca

Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, 1649-028 Lisboa, Portugal

[§] *Present adress: Instituto de Investigação e Inovação em Saúde, Porto, Portugal*

Author contribution

Rita Vaz-Drago, Noélia Custódio and Maria Carmo-Fonseca designed the experiments. Rita Vaz-Drago performed cell culture, biochemichal assays, image acquisition and analysis. Ana C de Jesus performed cell culture, image acquisition and analysis. Robert M Martin and Célia Carvalho generated the cell lines used in this study. José Rino contributed with expertise and helped in image analysis.

3.4.1. Overview

As in splicing, deregulation 3' end processing can be the cause of many human disorders. Thus, studying the kinetics of pre-mRNA cleavage and transcription termination is essential to understand the molecular mechanisms that may be disrupted in the context of human disease.

Compared to capping, splicing, transcription initiation and elongation, the last steps of mRNA biogenesis have been less studied. Here, we directly examined with single-molecule sensitivity the timing of pre-mRNA cleavage/polyadenylation and termination in the nucleus of living human cells. Using reporter transcripts labelled with MS2 or PP7 stem loops inserted upstream of the poly(A) site, we show that it takes 15-30 seconds to cleave and release the fully transcribed nascent RNA from the site of transcription. As expected, escape of the newly synthesized mRNA from the site of transcription is significantly delayed upon knocking down the essential cleavage and polyadenylation factor CPSF3. Analysis of reporter transcripts with λN stem loops inserted downstream of the poly(A) site reveals that these RNAs are also released from the site of transcription within 30 seconds after synthesis. Taken together, these results indicate that key steps in mRNA biogenesis including cleavage and polyadenylation can occur in just a few seconds, which is much faster than previously thought.

3.4.2. β-globin pre-mRNA molecules with stem loops inserted in exon 3 are efficiently spliced and cleaved

Synthesis of mRNA in mammalian cells comprises several processes including transcription initiation and elongation, splicing, cleavage, polyadenylation, release and termination of the nascent transcript. Kinetics of transcription initiation and elongation as well as the timing of pre-mRNA splicing have been previously estimated (Martin, Rino et al. 2013). However, 3' end processing and transcription termination kinetics have been less studied. To determine, with single molecule resolution the time of release of RNA molecules in living cells, we inserted binding sites for the coat protein of bacteriophage MS2 in the terminal region of the β-globin gene (*HBB*), either in the second (β-M2) or in the last exon (β-M3).

The *HBB* transcript is composed of three exons and two constitutively spliced introns, and the molecular mechanisms that lead to cleavage and 3' end processing have been extensively studied using biochemical methods (West, Proudfoot et al. 2008).

A single copy of β-M2, β-M3 or a control transgene without stem-loop sequences (β-WTΔ) was stably integrated into the genome of Flp-In T-Rex-293 cells, through site-specific DNA recombination, under inducible CMV promoter control (Tet-On Expression System) (**Figure 18A**). To perform live-cell visualization of transcripts, the above described cell lines were transiently transfected with a plasmid encoding a fusion protein comprising a GFP linked in-frame to the carboxyl terminus of MS2 coat protein (MS2-GFP). A nuclear localization signal was inserted in the MS2-GFP construct to favor binding of this fusion protein to the nascent RNA molecules.

The MS2 recognizing sequence is composed by a 4 nt loop and a 7 bp stem that harbors a single adenine bulge, forming a 19 nt hairpin structure. Each hairpin is coated by two MS2-GFP proteins which form a dimer before binding. Because the fluorescence emitted by one MS2-GFP dimer is not sufficient to be detected within living cells using our imaging system, 24 stem-loops sequences were inserted in the mentioned gene regions (Urbanek, Galka-Marciniak et al. 2014).

To determine whether the insertion of MS2 binding sites in the β-globin coding regions interfered with RNA processing, we carried out RT-PCR and RT-qPCR for splicing pattern and cleavage analysis. Total and chromatin-associated RNA were isolated from cells transfected and induced in the same conditions used for live cell microscopy. Splicing pattern and amount of uncleaved transcripts of β-M2 and β-M3 were compared with a cell line containing an untagged version of the β-globin gene (β-WTΔ), using the primer pairs indicated (**Figure 18B, Table M3**). The results indicated that the efficiency of splicing and cleavage are not affected when the MS2 loop sequences are inserted in the last exon of the β-globin gene (β-M3). However, when this 1176bp MS2 sequence is inserted in the middle of the gene body (exon 2), we detected an increase in the amount of unspliced products, for both first and second introns. We also detected an increase in uncleaved transcripts in this cell line, when compared to the β-WTΔ and β-M3. Thus, comparing with an untagged version of the HBB transcript, the β-M3 transcript does not present any splicing and cleavage defects and can be used in the determination of the time of cleavage of the HBB transcript.

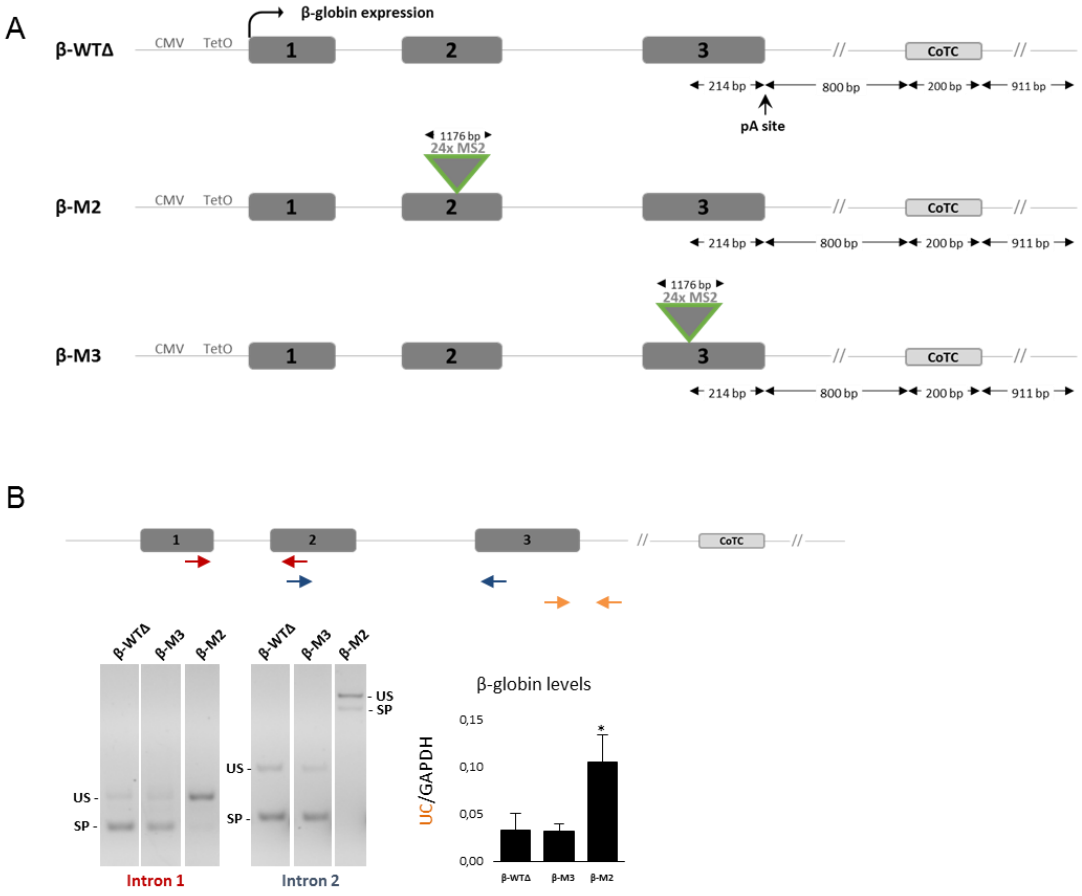


Figure 18. Biochemical analysis of splicing and cleavage efficiencies for the HBB pre-mRNA molecules.

A) Schematic illustration of the structure of three HBB transgenes used in this study. Tetracycline-inducible expression is under the control of a human CMV promoter. Binding sites for MS2 were inserted into either the second or third exon. B) RT-PCR of HBB RNA unspliced (US) and spliced (S) and RT-qPCR analysis of *HBB* RNA uncleaved (UC). Total and chromatin-associated RNA was extracted and reverse amplified using random primers. RT-PCR and RT-qPCR primers to detect different HBB isoforms are represented as coloured arrows. The asterisk denotes statistically significant differences (Student's *t*-test $*p<0.05$).

3.4.3. The time of release of β-globin transcripts from the transcription site ranges between 15 and 25 seconds.

Cells were imaged using a spinning-disk confocal microscope after the expression of the single-copy reporter gene β-M3 with doxycycline for 4 hours. Images were recorded every 5 seconds, for a variable number of timepoints (frames), but no longer than 80 timepoints (~6 minutes) per cell. Each timepoint analysed consists of the highest-intensity z-plane at 0,27 µm thickness of a total of 8 imaged z-planes. RNAs were imaged as diffraction-limited particles that were than analysed using the STaQTool program (Rino, de Jesus et al. 2016) (**Figure 19**). Before analyzing the fluorescence variation at the transcription site (TS), a calibration step was performed in which a fluorescence intensity (TFI) range and a width (W) range were determined for single fluorescent transcripts. Since mRNA molecules diffuse individually throughout the nucleus (Fusco, Accornero et al. 2003, Shav-Tal, Darzacq et al. 2004), nucleoplasmic-released mRNAs correspond to a unique population and the range of TFI and W values were calculated using the image analysis program STaQTool (Rino, de Jesus et al. 2016) (**Figure 19)**. Following this calibration step, TFI values can be translated into the average number of RNA molecules present at the site of transcription. After the identification of the transcription site, we performed single spot tracking on nascent RNAs and determined the TFI of single RNA molecules for each time point of acquisition. Cycles of fluorescence increase and loss corresponding to the permanence of the fluorescently labelled exon 3 of β-M3 transcript were then manually identified.

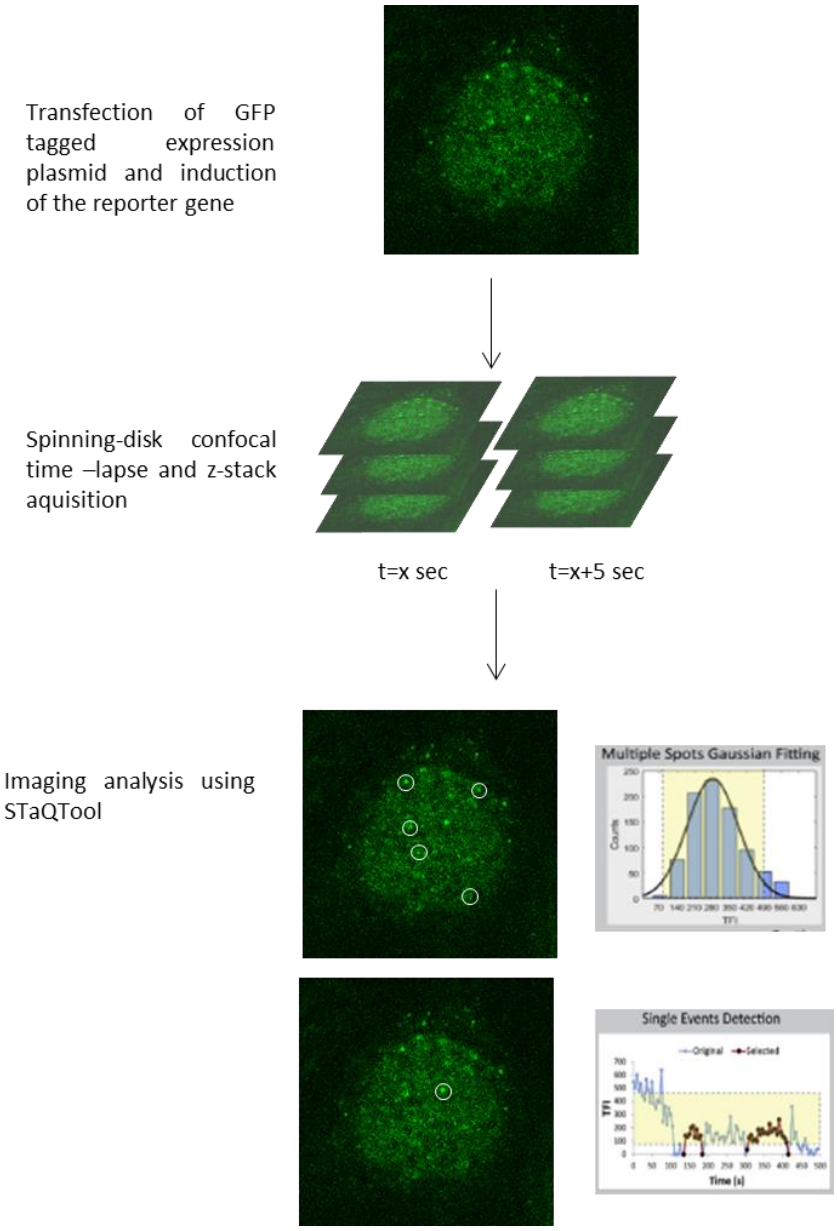


Figure 19. Identification of the transcription site and nucleoplasmic released RNA molecules.

A spinning-disk confocal microscope was used to obtain lime-lapse and z-stack images of cells expressing β-M3 transcripts tagged with MS2-GFP in the third exon. z stacks of optical sections were obtained every 5 seconds (sec). Maximum-intensity projection images of fluorescence at the transcription site were generated for each time point and pseudocolored. The software identifies diffraction-limited spots in the cell nucleus at each time point and performs Gaussian fitting on the TFI and W distributions and outputs their mean values as well as upper and lower values for the 68% (1 standard deviation) and 95% (2 standard deviations) range. Additionally, it plots the TFI at the transcription site over time for the construct analysed. Adapted from (Rino, de Jesus et al. 2016).

As the RNAPII transcribes the last exon of the *HBB* reporter gene containing the 24 repeated hairpin sequence, fluorescently coated MS2 proteins bind as homodimers to the newly synthesized RNA hairpin structures. The cumulative binding of MS2-tagged proteins to the newly synthesized RNA molecule over time leads to an increase in fluorescence until the RNA stem-loops are fully synthesized and a maximum fluorescence intensity is reached. It is then expected that RNAPII proceeds through the remaining exonic sequence (80 bp coding and 134 bp 3' UTR sequences) and downstream 3' sequence. During this time, MS2-GFP-tagged nascent transcript remains attached to RNAPII at the transcription site and fluorescence intensity is maintained, until cleavage and release of the nascent transcript, corresponding to a decrease in fluorescence intensity (**Figure 20 A** and **B**).

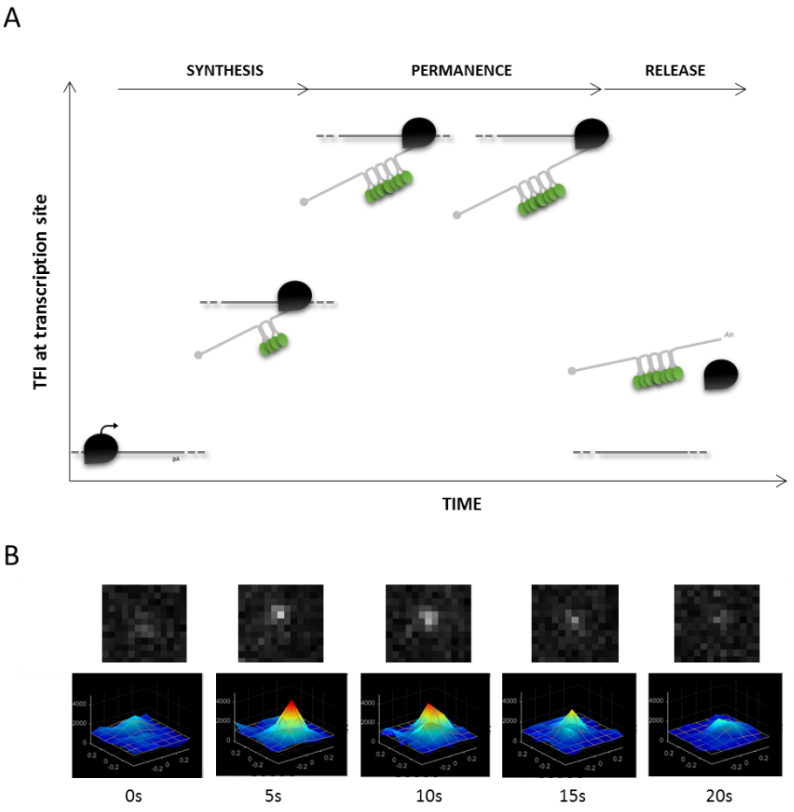


Figure 20. Detection of fluorescent cycles corresponding to synthesis of a single pre-mRNA molecule.

A) Diagram of expected single nascent transcript fluorescence intensity trajectory over time. B) The interaction of MS2 proteins fused to GFP with the exonic stem loops allows the transcribed pre-mRNAs to be visualized over time and cycles of synthesis-permanence-release to be identified.

With this system we were not able to determine exactly when cleavage, 3' end processing and release of the nascent transcript occur within the fluorescence cycles observed: it may occur during maintenance and/or decrease of fluorescence signal. Thus, to avoid the introducing an analysis bias we considered time of cleavage/release as the time of the fluorescence cycle duration, which includes the time of synthesis, permanence and release of the nascent transcript.

We manually identified 64 synthesis-permanence-release cycles (events) based on fluorescence fluctuations and independently of RNAPII occupancy. Among those 64 events, 22 synthesis-permanence-release cycles were immediately preceded by background levels of fluorescence, indicating we were measuring single-molecule events (**Figure 21A**).

Furthermore, we hypothesized that if the observed fluorescence cycle duration results from efficient RNA transcription and processing, knocking down the 3' end processing endonuclease CPSF3 will lead to an increase in cycle duration. CPSF73 was shown to be the endonuclease that cleaves the nascent RNA at the CA dinucleotide (Mandel, Kaneko et al. 2006, Shi and Manley 2015). Thus, we performed time-lapse and multiplane imaging of nascent transcripts in single cells upon disruption of 3' end processing mechanism and evaluated the effect of this interference based on cycle duration. We manually identified 54 synthesis-permanence-release events, 29 of those correspond to single RNA molecules visualized at a time. Upon CPSF3 KD, the time of a cleavage of β-globin transcript at the TS increases to 20-40 secs (**Figure 21B** and **C**).

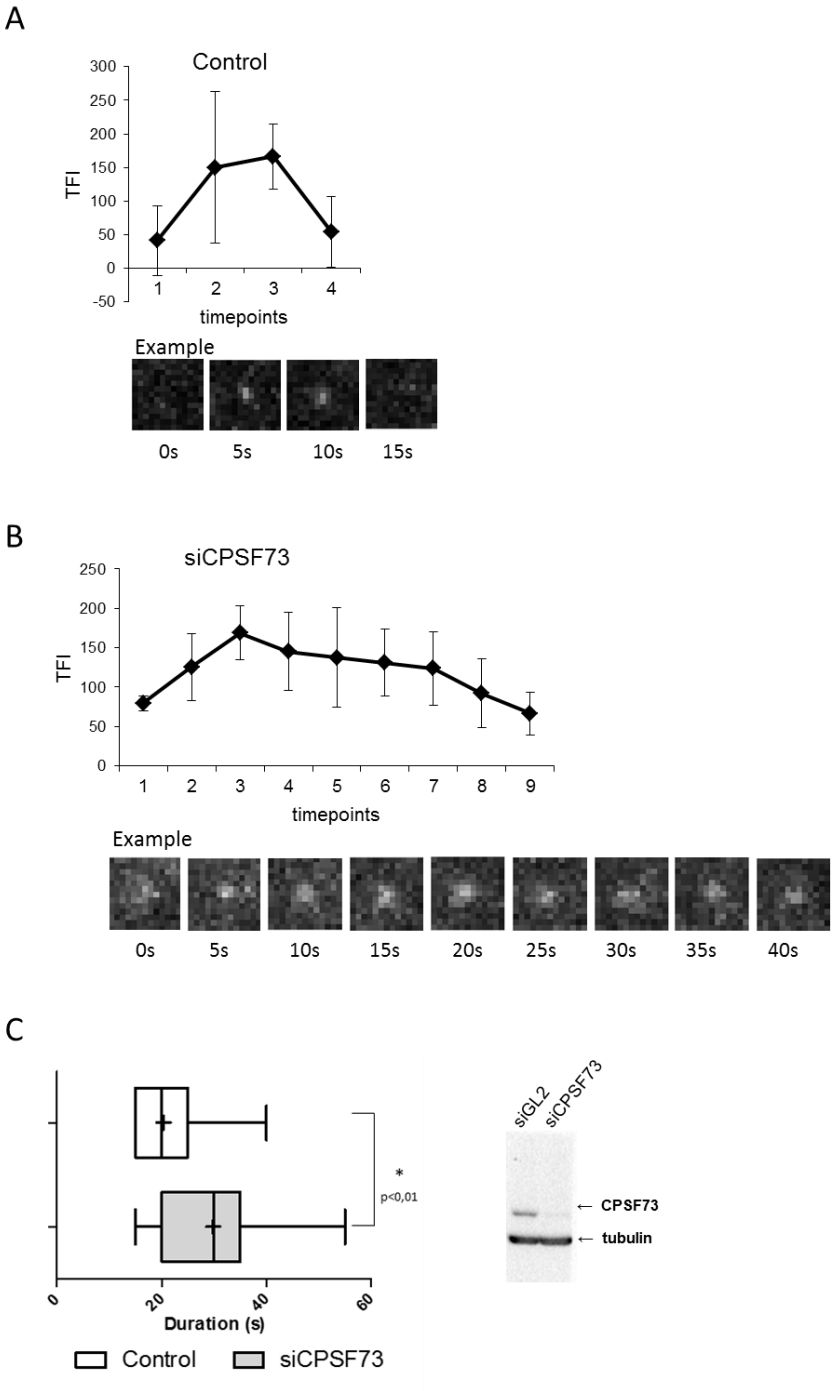


Figure 21. Identification of fluorescent cycles corresponding to synthesis, permanence and release of a single *HBB* pre-mRNA molecule.

A and B) Average TFI fluctuation of single-molecule events at TS over time in A) control cells and B) CPSF3 depleted. Each plot (Control and siCPSF3) is accompanied by a representative example of a synthesis-permanence-release event. C) Left panel: box plot representing the distribution of the duration of cycles of synthesis-permanence-release in Control and CPSF3 depleted cells. The asterisk denotes statistically significant differences (Kruskal-Wallis, $*p<0.05$, $**p<0.01$). Right panel: Western blot showing knockdown efficiencies of siRNA treatments for CPSF73.

3.4.4. IgM pre-mRNA molecules with stem loops inserted in the last exon are efficiently spliced, cleaved and terminated

Having determined the time of release of single *HBB* transcripts, we were next interested in determining whether the time of cleavage obtained for the β-M3 was specific of this construct or whether it represents a more general feature of other genes. We then focused our analysis on the mouse immunoglobulin μ reporter gene, here referred as IgM minigene.

In this case, we took advantage of the PP7 system, previously introduced by Larson et al, that is based on a similar principle as the MS2 system (Larson, Zenklusen et al. 2011). We inserted in the last exon of the IgM minigene a sequence with 24 stem-loop sequences containing a 6 nt loop and an 8 bp stem loop, that has a high-affinity interaction with a phage protein.

Furthermore, we were interested in determining how cleavage/release is interconnected with splicing and termination of transcription. For that, MS2 binding sites were inserted in the intron and λN22 binding sites were inserted after the poly(A) and cleavage site of the IgM minigene (**Figure 22A**). The λN22 system is the second most frequently used RNA imaging system after the MS2 system and is based on the recognition of a 15 nt hairpin constituted by a 5 nt loop and 5 bp stem recognized by the N phage protein. A single copy of this triple-labelled transcript was stably integrated into the genome of Flp-In T-Rex-293 cells, through site-specific DNA recombination, under inducible CMV promoter control and Tet-On Expression System, generating the IgM-MPB cell line (**Figure 22A**).

To determine whether the insertion of MS2, PP7 and λN22 binding sites in the IgM minigene interfered with RNA processing, we carried out RT-qPCR for analysis of splicing, cleavage and termination efficiencies. As a control and for comparison, we generated a cell line expressing an IgM reporter gene that does not contain stem-loop sequences at the poly(A) and past poly(A) regions, herein called IgM-MS2 (**Figure 22A**). We purified RNA from the chromatin, nucleoplasm and cytoplasm fractions and performed RT-qPCR to detect spliced and unspliced isoforms (**Table M3**). The ratio unspliced/spliced (US/SP) isoforms decreased from the chromatin to the nucleoplasm fraction, being almost zero in the cytoplasm, which indicates that this transcript is spliced

efficiently in the nucleus (**Figure 22B**). Simultaneously, a cyan fluorescent protein (CFP) coding sequence fused to a peroxisomal targeting sequence (PTS) was inserted in-frame in the second exon of the IgM minigene, allowing us to confirm that this engineered transcript was correctly spliced and exported to the cytoplasm. Consistent with the RT-qPCR result, cyan fluorescence was detected in the cytoplasmic peroxisomes (**Figure 22B**). Additionally, RNA purified from the chromatin fraction was reverse transcribed and used as template to determine the amount of uncleaved (UC) and unterminated (UT) IgM transcripts. UC and UT RT-qPCR values were normalized to the CPF-PTS coding region. The results show that the amounts of UC and UT transcripts in the IgM-MPB cell line are lower compared to the values obtained in the IgM-MS2 cell line that does not contain stem-loops in the 3' region of the gene. This shows that the efficiencies of cleavage of the IgM transcript and termination of transcription are not perturbed by the insertion of PP7 and λN22 binding sites.

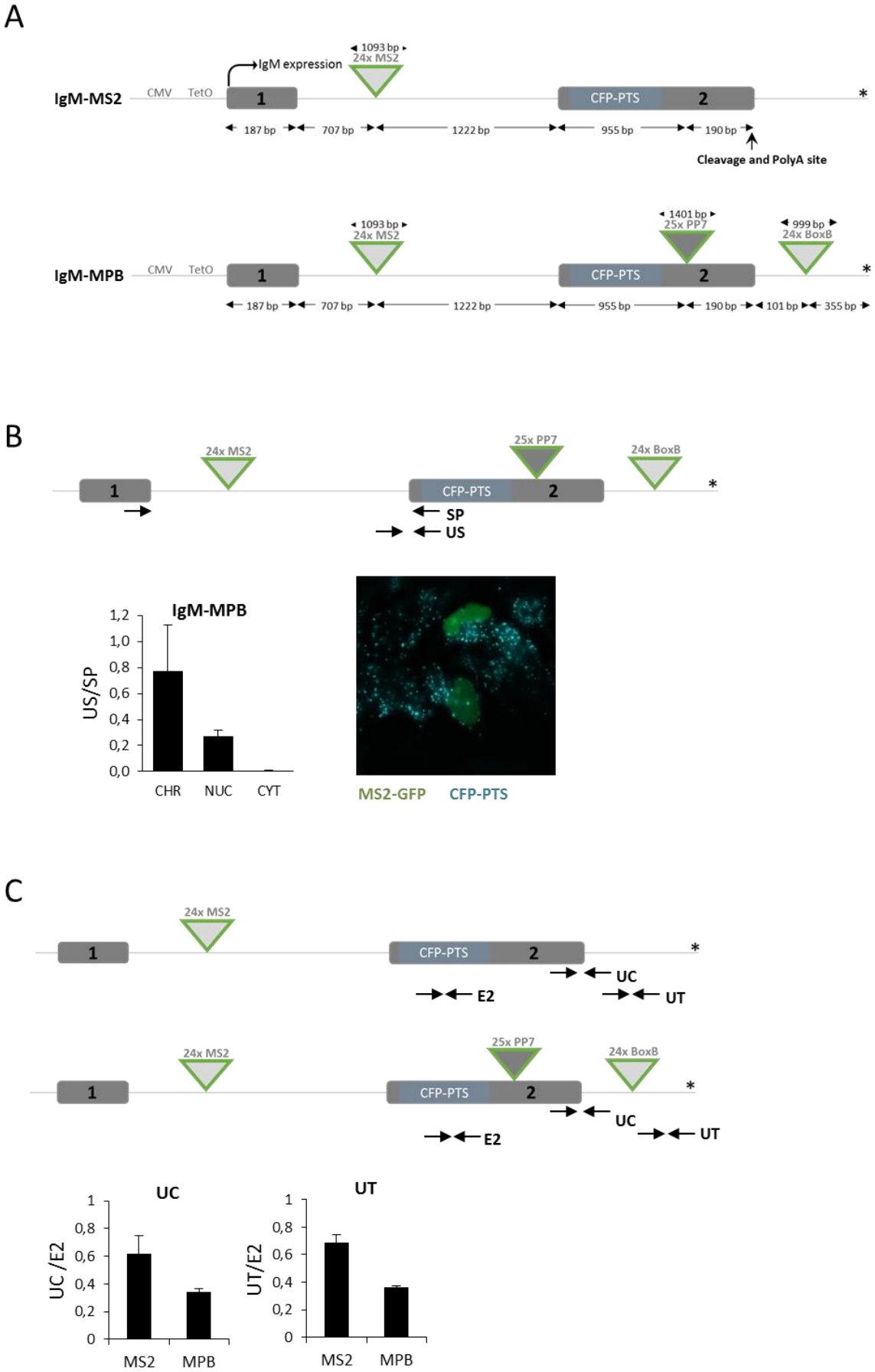

A
IgM-MS2
CMV
TetO
IgM expression
1093 bp
24x MS2
1
CFP-PTS
2
187 bp
707 bp
1222 bp
955 bp
190 bp
Cleavage and PolyA site
IgM-MPB
1401 bp
25x PP7
999 bp
24x BoxB
101 bp
355 bp
B
SP
US
IgM-MPB
US/SP
1,2
1,0
0,8
0,6
0,4
0,2
0,0
CHR
NUC
CYT
MS2-GFP
CFP-PTS
C
UC
E2
UT
UC /E2
UT/E2
MS2
MPB

Figure 22. Biochemical and image analysis of splicing and cleavage efficiencies for the IgM pre-mRNA molecules.

A) Schematic illustration of the structure of two IgM transgenes used in this study. Tetracycline-inducible expression is under the control of a human CMV promoter. Binding sites for MS2 were inserted into the intron, binding sites for PP7 were inserted in the second exon and binding sites for λN22 were inserted in the past-polyA region of the IgM minigene. B) Analysis of splicing efficiency of IgM-MPB by RT-qPCR in which ratio of unspliced/spliced (US/SP) was calculated, and by the visualization of the protein product by confocal microscopy. C) RT-qPCR analysis of IgM-MPB RNA uncleaved (UC) and unterminated (UT) isoforms normalized to the expression level of exon 2. Chromatin-associated RNA was extracted from and reverse amplified using random primers. RT-PCR and RT-qPCR primers to detect different IgM isoforms are represented as black arrows. * represents recombination site.

3.4.5. β-globin and IgM transcripts take similar time to be released from the transcription site

In accordance with the torpedo model of termination, cleavage may regulate the timing of release and polyadenylation of nascent transcripts but may also influence transcription termination (Proudfoot 2016).

β-globin was identified as having a co-transcriptionally cleaved (CoTC) type of termination element (Dye and Proudfoot 2001), while transcription of the IgM reporter gene terminates via a mechanism that is independent of a CoTC element. In CoTC-type of termination, cleavage first occurs at an A/T-rich element 1-2kb downstream of the poly(A) site, allowing XRN2 entry and degradation of the downstream RNA molecule.

Single-molecule kinetics analysis of these two types of cleavage has never been performed before, thus, the analysis of β-globin and IgM might therefore allow investigation of the similarities and differences in the timing of CoTC-dependent and CoTC-independent release/transcription termination.

To do that, we performed single molecule analysis and compared the time of diffusion of β-M3 RNAs away from the transcription site with the time of diffusion of IgM-MPB RNAs. After PP7-GFP plasmid transient transfection and reporter gene expression induction, we carried out time-lapse and multiplane imaging using a spinning-disk confocal microscope. We performed single spot

tracking on nascent RNAs after identification of the transcription site. Using the STaQTool program we determined the fluorescence intensity of single RNA molecules for each timepoint of acquisition. Based on that, we manually identified 64 synthesis-permanence-release cycles (events) based on fluorescence fluctuations. Finally, we graphically plotted the distribution of cycle duration. Among those, 29 synthesis-permanence-release cycles were immediately preceded by background levels of fluorescence (**Figure 23**). Interestingly, time of release of the IgM transcript from the TS is 15-25 secs, similarly to what has been previously described for the HBB construct. From this, we conclude that different transcripts take the same time to be released from the transcription site independently of the presence of the downstream cleavage element CoTC.

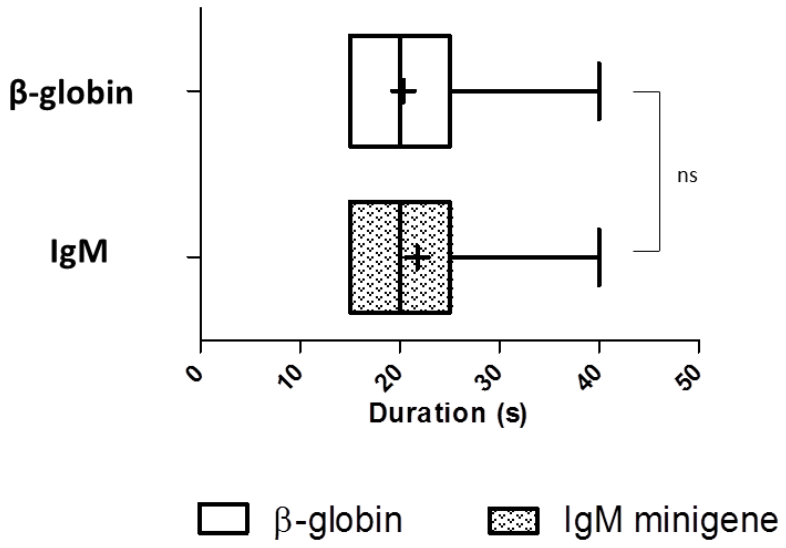


Figure 23. Distribution of time of release for HBB and IgM transcripts.
Box plot displaying the distribution of the duration of cycles of synthesis-permanence-release in β-globin (β-M3) and IgM (IgM-MPB).

3.4.6. Different processing steps have different kinetics

Time-based studies utilizing live-cell imaging have been previously used for the analysis of transcription elongation and splicing kinetics (Schmidt, Basyuk et al. 2011, Martin, Rino et al. 2013, Coulon, Ferguson et al. 2014, Tantale, Mueller et al. 2016). However, comparatively to other processes of the transcription cycle, transcription termination has been the most understudied (Darzacq, Shav-Tal et al. 2007, Palangat and Larson 2016). To overcome this gap, we decided to directly measure the time of termination of transcription of the previously used IgM reporter by live-cell microscopy. To make the IgM RNA visible at the past poly(A) region, we inserted BoxB stem-loops that are bound by λN fluorescent proteins once the RNAPII transcribes the downstream region of the cleavage site. Complementarily, we decided to determine the time of splicing of IgM pre-mRNA by inserting MS2 binding sites in the intronic region of the transcript. After MS2-GFP plasmid transfection, we measured fluctuations in MS2 fluorescence intensity at the transcription site which allowed estimation of intron lifetime as a measure of time of splicing. With this approach, we were able to measure kinetics splicing and termination for the same reporter.

By looking at the λN-GFP fluorescence signals at the transcription site, we were able to determine that transcription took around 20 to 80 seconds to terminate after the ribonucleotide synthesis of the past poly(A) region (**Figure 24**). Consistently with what has been previously described for β-globin and IgM transgenes (Martin, Rino et al. 2013), we determined that time of splicing, assessed as a measure of intron lifetime, ranges between 20 and 50 seconds. For comparison, the time distribution of release previously determined is also indicated (**Figure 24**).

In summary, our results showed that different transcripts are processed with similar kinetics, but different processing mechanisms occur in a specific time window.

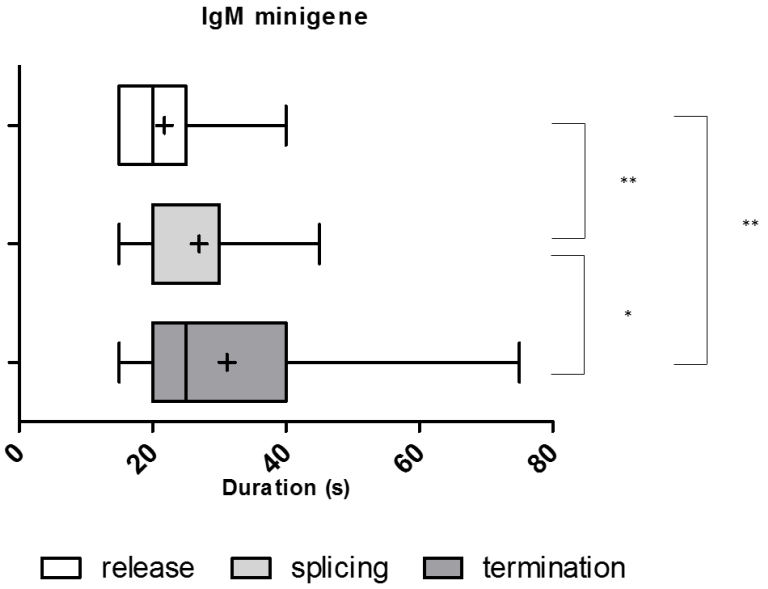


Figure 24. Distribution of time of release, splicing and termination of IgM transcripts. Box plot displaying the distribution of the duration of cycles of synthesis-permanence-release of different regions throughout the pre-mRNA of IgM-MPB (exon, intron and past-poly(A) region). The asterisk denotes statistically significant differences (Kruskal-Wallis, *p<0.05, **p<0.01)

3.4.7. CPSF73 KD specifically delays time of release

Molecular interactions between mRNA processing reactions and transcription have been extensively described [reviewed in (Proudfoot, Furger et al. 2002, Bentley 2014)]. Specifically, connections between 3' end processing and splicing have been demonstrated. Most of those conclusions were based on studies where mutations in poly(A) signals were shown to reduce the efficiency of splicing of the upstream intron (Rigo and Martinson 2008). However, protein-protein interactions between cleavage, polyadenylation and splicing factors were also demonstrated to play a role in exon definition of the last exon (Kyburz, Friedlein et al. 2006). Signals that regulate 3' end processing were also shown to play a role in transcription termination, since efficient termination requires a functional poly(A)signal. Additionally, it was shown that depletion of CPSF73 substantially reduced RNAPII pausing over the cleavage site, reflecting transcriptional termination defects (Nojima, Gomes et al. 2015).

To establish the impact of 3' end processing on the kinetics of release, splicing and transcription termination by live-cell microscopy, we depleted CPSF73 by RNAi and measured fluorescence cycles at the transcription site of IgM

minigene. We start by measuring synthesis-permanence-release fluorescence cycles of the 3' region of the transcript and found that IgM-MPB transcripts take longer to leave the chromatin template upon downregulation of CPSF73, which is in accordance with what we had observed for the β-M3 transcript (**Figures 21A and 25**). In contrast, we found that CPSF73 depleted cells showed no splicing or termination defects, as intron and past poly(A) transcript lifetimes were not significantly changed when compared with control cells expressing normal levels of the cleavage and polyadenylation factor (**Figure 25**). Thus, our data showed that CPSF73 knock-down specifically delays time of release of IgM MPB transcripts, which argues against the coupling model proposed for this transcript.

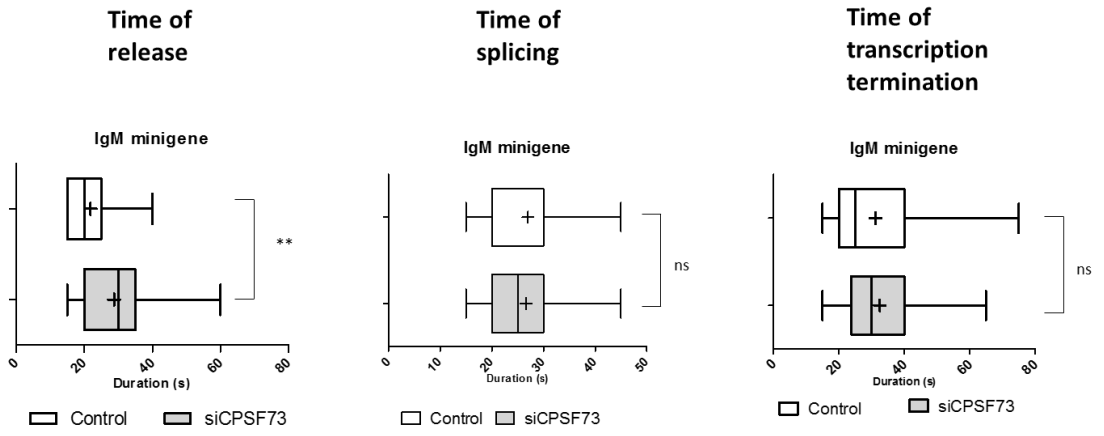


Figure 25. Distribution of time of release, splicing and termination of IgM transcripts upon CPSF73 KD.

Box plot displaying the distribution of the time of release, splicing and transcription termination in control and CPSF73 KD cells. The asterisk denotes statistically significant differences (Kruskal-Wallis, $^{**}p<0.01$)

4. Discussion

Deregulation of pre-mRNA processing is one of the main causes of genetic disease in human individuals. The main goal of this work was to elucidate the impact that defects in pre-mRNA splicing and cleavage have on mRNA biogenesis, stability and kinetics. The systems we used in these studies consisted of patient-derived cell lines (both wild-type and several splicing mutants) and engineered HEK 293 cell lines with inducible expression of a variety of versions of two reporter genes. The main discoveries of this work were that splice site mutations trigger chromatin-associated RNA surveillance responses that contribute to down-regulate the expression of abnormal mRNAs (Chapter 3.1), that deep-intronic mutations most often create donor splice sites that lead to the inclusion of pseudo-exon in the mRNA (Chapter 3.2), and that release of reporter transcripts from the transcription site occurs within seconds and with specific kinetics that is distinct from splicing and transcription termination (Chapter 3.4). We analysed six cell lines derived from patients carrying splice-site mutations and in three of them we found reduced mutant RNA levels associated with chromatin. In two of these lines, lower abundance of mutant chromatin-associated RNA correlated with reduced transcriptional activity. Additionally, we analysed published data from patients carrying deep-intronic splicing mutations and showed that pseudo-exon inclusion is the main consequence for mRNA biogenesis caused by the creation and activation of deep-intronic splice sites. Moreover, we highlight the importance of selecting the appropriate nascent RNA purification protocol in order to obtain the most accurate calculation of co-transcriptional splicing efficiency (Chapter 3.3). Regarding the study of 3' end kinetics, we found that release of a fully transcribed single mRNA molecule from the site of transcription is accomplished in 15-25 seconds and it is specifically regulated by the activity of a cleavage and polyadenylation factor with known relevance in the context of human disease.

A co-transcriptional quality control operates in patient-derived cell lines

A link between splicing mutations and transcription has been previously described (Kwek, Murphy et al. 2002, Damgaard, Kahns et al. 2008). In the

study by Damgaard et al, mutations in the promoter-proximal 5' splice site were shown to severely decrease transcription by a mechanism that involved U1 snRNA recognition and assembly of the preinitiation complex (Damgaard, Kahns et al. 2008). Here we observe a similar scenario for the *TAZ* 5' splice site mutant (SM). However, we also detected decreased transcription of the *MARVELD2* 3'SM gene, which contains a 3' splice site mutation in the third intron. This observation raises the possibility that additional mechanisms are involved in coupling transcription to splicing efficiency. Indeed, inefficient splicing can cause stalling of spliceosomes on the transcripts, leading to recruitment of the RNAi machinery, heterochromatin formation and transcriptional silencing (Bayne, Portoso et al. 2008, Dumesic, Natarajan et al. 2013). Down-regulating the transcription of mutant genes appears 'economical', as it saves energy in producing and discarding aberrant RNAs. Yet, many transcripts produced from genes with splicing mutations escape this type of control.

Although transcription from the *TAZ* 3'SM and *XPC* 3'SM genes is similar to wild-type, RNA levels associated with chromatin differ significantly. The steady-state level of chromatin-associated *TAZ* 3'SM transcripts is higher than wild-type, whereas *XPC* 3'SM transcripts are reduced compared to wild-type. The results obtained with *TAZ* 3'SM transcripts is reminiscent of our previous observations with β-globin splicing mutants (Custodio, Carmo-Fonseca et al. 1999, de Almeida, Garcia-Sacristan et al. 2010), suggesting that abnormally processed RNAs persist associated with the chromatin template and consequently accumulate in this fraction. In contrast, the results obtained with *XPC* 3'SM suggest that these transcripts undergo a fast co-transcriptional decay most likely mediated by the exosome and/or XRN2 (Davidson, Kerr et al. 2012).

A main conclusion from this study is that disease-causing splicing mutations can have a variety of effects on mRNA biogenesis. For all disease-associated genes analysed, a single splice site mutation leads to expression of multiple mRNA isoforms. Some of these isoforms may contain a PTC due to a frame shift caused by activation of a cryptic splice site or exon skipping, others may be recognized as abnormally spliced due to intron retention, while others may not be recognized as faulty (namely, if the reading frame is not disrupted).

Thus, depending on the isoform expressed, the mutant RNAs may be differentially recognized by distinct surveillance mechanisms. We also found that *TAZ*, *MARVELD2* and *XPC* genes are expressed at low levels in immortalized lymphoblastoid cells. Since the proteins encoded by these genes have tissue-specific functions, it remains to be established whether the patterns of mRNA biogenesis observed in lymphoblastoid cells are physiologically representative. Another limitation of working with immortalized lymphoblastoid cell lines is that these cells were resistant to RNA interference manipulations aimed at identifying the nucleases responsible for mutant RNA degradation in the nucleus. For future studies, iPSCs derived from patients are likely to represent improved disease models. Differentiation of iPSCs into the specific cell types that require expression of the mutant genes for their normal function will provide a valuable system to address how cytoplasmic and nuclear quality control mechanisms operate to reduce expression of abnormal RNAs caused by splicing mutations.

Deep-intronic variations: a source of disease causing-mutations

Despite major advances in clinical genetic analysis introduced by the application of next-generation sequencing strategies, approximately half of the patients remain without a precise genetic diagnosis, which represents a significant limitation for clinical care. Here we highlight that DNA intronic variants located throughout introns can be the cause of human disease and should be investigated when first line approaches such as next-generation sequencing-based gene panels, whole-exome sequencing, microarray and multiplex ligation-dependent probe amplification-based deletion/duplication analysis fail to identify a causative mutation.

In order to find novel deep intronic mutations and determine their pathogenicity it is crucial to combine sequencing of intronic regions with studies addressing the mRNA molecules produced in affected tissue from patients. This can be done by conventional RT-PCR analysis and sequencing of cDNA products, or by direct RNA-seq analysis. Examination of the patient's transcriptome has the advantage of rapidly detecting the presence of abnormal

splicing isoforms (Gonorazky, Liang et al. 2016). Reduced levels of mutant transcripts are normally indicative of a disease-causing mutation that either disrupts normal splicing and targets abnormal mRNAs for degradation or inactivates a transcriptional regulatory motif. However, in some cases the mutation interferes with regulatory motifs or non-coding RNAs that control the expression of other genes. RNA-seq is clearly the best approach for quickly identifying these situations.

It is now clear that introns harbor a multiplicity of functional or potentially functional elements and that variants located deeply within introns can lead to human disease through a variety of molecular mechanisms: creation and activation of non-canonical splice sites (ss), activation of splicing enhancer motifs, abolishment of splicing silencer motifs and transcription enhancer elements and disruption of non-coding RNA gene sequences.

Deep intronic mutations that contribute the most to genetic disease tend to affect splicing-related mechanisms and, lead to the inclusion of pseudo-exon in the mRNA. Less frequently, these types of mutations cause retention of intronic sequences through weakening of canonical ss. To our knowledge, exclusion of authentic exonic sequences is an extremely rare consequence of deep intronic mutations and was only reported once for a mutation in the last intron of the *GBE1* gene (Akman, Kakhlon et al. 2015).

Pseudo-exon inclusion in the mRNA molecule most often result from the creation of a new donor ss by mutation followed by the activation of a pre-existing upstream non-canonical acceptor ss. Interestingly, the majority of pathogenic mutations responsible for the generation of new boundaries in authentic exons (activation cryptic ss) are located within canonical donor ss rather than acceptor ss (Vorechovsky 2006, Buratti, Chivers et al. 2007, Buratti, Chivers et al. 2011). Moreover, it has been proposed that the correct recognition of a donor ss drives correct recognition of an upstream acceptor ss (Chen, Huo et al. 2000, Krawczak, Thomas et al. 2007) and both contribute to exon definition [reviewed in (De Conti, Baralle et al. 2013, Hollander, Naftelberg et al. 2016)]. Thus, upon activation, pseudo-exons seem to rely on the same mechanisms that allow authentic exons to be included in the mRNA. Remarkably, compared to authentic exons, pseudo-exons tend to be smaller in size. Interestingly, it was suggested that alternatively spliced exons have

weaker splicing signals and shorter exon length (Zheng, Kwon et al. 2005). Recognition of exons is regulated by the presence of recognizable ss but also *cis*-acting splicing regulatory elements. The reduced presence or complete absence of those elements may favor the inclusion of pseudo-exons with lower nucleotide content (Grellscheid and Smith 2006).

In addition to *cis*-acting elements the secondary structure of the pre-mRNA is increasingly recognized as a general modifier of splicing. In accordance with that, perturbation of the RNA structure has also been proposed as a consequence of deep intronic mutations. A point mutation located within intron 5 of the *CLASP1* gene was described as pinpointing a modification of the secondary structure of the *U4ATAC* snRNA gene, which would further compromise its interaction with other spliceosomal components, leading to general splicing defects of U12-dependent introns (Edery, Marcaillou et al. 2011). Structural intronic features were also proposed as key regulators in pathological *ATM* and *CFTR* pseudoexon inclusion events (Buratti, Dhir et al. 2007).

Furthermore, binding sites of regulatory proteins are other regions that are often disrupted upon deep intronic mutations. Most of them cause inclusion of pseudo-exon in the mRNA and decrease in mutant gene transcription due to changes in molecule affinity of RNA and DNA binding proteins, respectively. It remains to clarify if alterations in DNA methylation, histone marks, nucleosome positioning and the kinetics of transcription are associated with this type of mutation. If so, screening of chromatin marks would give a good indication of the pathogenic consequence of a deep intronic variant and how it may trigger QC mechanisms, similarly to splice-site mutations.

Purification of nascent transcripts may have an impact on the calculation of splicing efficiency

In mammalian cells, the majority of introns are removed co-transcriptionally, while pre-mRNA is attached to the chromatin by RNAPII (Han, Xiong et al.

2011, Schmidt, Basyuk et al. 2011, Brugiolo, Herzel et al. 2013). This implies not only a mechanistic coupling between the spliceosomal proteins and RNAPII, but also a temporal coupling, since co-transcriptional splicing depends on the rate of elongation and termination of the transcriptional machinery (Fong, Kim et al. 2014, Naftelberg, Schor et al. 2015, Aslanzadeh, Huang et al. 2018). Therefore, the isolation of nascent transcripts facilitates the analysis of co-transcriptional splicing efficiency, by enriching the sample in this sub-population that constitutes 5-10% of total cellular RNAs.

Nascent transcripts can be purified from the chromatin fraction or by a short-time of metabolic labelling with a uridine analogue. The tight interaction between RNAPII and the DNA template combined with the sedimentation properties of chromatin have been widely used to isolate nascent RNAs to study different splicing properties, including its coupling to transcription (Pandya-Jones and Black 2009, Bhatt, Pandya-Jones et al. 2012, Herzel and Neugebauer 2015). Although the metabolic labelling of nascent transcripts with 4sU is a powerful method to assess the kinetics of RNA metabolism, the quantification of splicing using this method has been described only once (Windhager, Bonfert et al. 2012). 4sU is readily taken by the cells and rapidly incorporated into nascent transcripts. An intrinsic property of this population of transcripts is that it is heterogeneous, ranging from a few to thousands nucleotides in length, reflecting the size of the gene, the position of the transcribing RNAPII and the kinetics of intron removal.

It has been reported that the use of HPDP-biotin to purify 4sU-tagged transcripts may be inefficient, as it leads to a bias towards the purification of longer transcripts that have incorporated more 4sU molecules. (Duffy, Rutenberg-Schoenberg et al. 2015). We, therefore, hypothesized that the calculation of splicing efficiencies in a population of nascent transcripts captured using this method could be underestimated due a preference for longer, unspliced transcripts during the biotinylation reaction.

RT-qPCR has been widely used to measure splicing efficiency (Vandenbroucke, Vandesompele et al. 2001, Ivings, Towns et al. 2008, Singh and Padgett 2009, Carrillo Oesterreich, Preibisch et al. 2010, Aitken, Alexander et al. 2011). Here we used the same approach to measure co- and post- transcriptional splicing efficiencies of four introns. We chose introns that differ in size (from 240 to

13000 nt) and kinetics of removal from the pre-mRNA (U2-/U12-dependent) and that were transcribed from highly expressed protein-coding genes in HEK 293 cells. Changes in splicing efficiencies, reflected in changes in the proportion of spliced transcripts, were assessed when cells were exposed to different times of 4sU incubation, and different nascent RNA purification methods (chromatin, HPDP-biotin and MTS-biotin) were applied.

It has been reported that when 4sU is used at concentrations ranging between 50 μM and 500 μM, for 48 hours, it can cause harmful effects on cell viability (Tani and Akimitsu 2012) and when used at final concentrations of 100 μM for 6 hours it can cause nucleolar stress and inhibit rRNA synthesis (Burger, Muhl et al. 2013). In our study, we performed ultra-short (2 minutes) and short (60 minutes) 4sU-tagging, and although we could not exclude that 4sU might induce some harmful effects in cells, we concluded that its incorporation into nascent transcripts does not affect splicing of the analysed introns.

Next, we compared the proportions of spliced transcripts in a population of 4sU-tagged transcripts purified either with HPDP- or MTS-biotin. The spliced ratios of all the introns analysed were significantly increased in the population of transcripts purified with HPDP-biotin relative to the population of MTS-biotin purified transcripts. Thus, the utilization of MTS-biotin substantially reduces the length bias associated with the purification of 4sU-tagged unspliced transcripts.

Taking together, our results show similar splicing proportions of nascent transcripts that were either unlabelled or tagged with 4sU, suggesting that this metabolic labelling procedure does not significantly interfere with the splicing reaction. Moreover, our data reveal that recovery of 4sU-tagged transcripts with HPDP-biotin is prone to bias towards the purification of longer unspliced pre-mRNAs that have incorporated more 4sU molecules, and thiol-mediated purification of nascent RNAs using MTS-biotin clearly reduces this bias. Altogether our results underscore the risk for bias in splicing kinetics analysis based on 4sU metabolic labelling, which may lead to underestimation of splicing efficiencies in populations of nascent transcripts with long introns.

Release of nascent transcripts from the transcription site is kinetically regulated

Live-cell and single-molecule studies have given a valuable insight that points towards precise kinetic regulation of transcription and pre-mRNA processing.

Many questions about the molecular processes that take place at 3' end of genes remain to be clarified. By using different RNA labelling methods, we show that release of nascent transcripts obeys to a precise timing that ranges from 15 to 25 seconds and that regulation of transcription termination seems be more kinetically permissive ranging between 20 and 80 seconds.

Previous live-cell imaging estimates suggested that the lifetime of a fully transcribed human β-globin RNA at the transcription site was around 116 seconds (Coulon, Ferguson et al. 2014). In a report from the same research group, it was estimated that transcription of an inducible exogenous chimeric ecdysone receptor gene took around 450 seconds to terminate (Palangat and Larson 2016).

The different analysis and the choice of the cell lines may contribute to the discrepancy among live-cell imaging studies in estimates of time of release and transcription termination. In our study we used HEK 293 cell and directly identified synthesis-permanence-release cycles of fluorescence that were immediately preceded by background levels of fluorescence as the behavior of single transcripts at the transcription site. In the other studies, U2OS cells were imaged and an auto-correlation function was applied to the fluorescence fluctuations of a bulk of RNA molecules being simultaneously transcribed.

Indeed, mammalian gene transcription appears to be dominated by bursting with periods of high transcriptional activity followed by periods of inactivity. (Dar, Razooky et al. 2012). In contrast to human promoters, viral promoters integrated in mammalian cells show constitutive expression (Yunger, Rosenfeld et al. 2010). This may be a problem when performing single-molecule kinetics analysis. To overcome this, we used low levels of doxycycline to induce both β-globin and IgM minigene expression, therefore increasing the number of cells synthesizing a single reporter transcript during image acquisition. Nevertheless, we cannot exclude that the widely used insertion of stem-loops

and the binding of coat proteins to the newly synthesized RNA molecules may perturb kinetics of processing and transcription termination.

Transcription termination is directly coupled to pre-mRNA processing. So far, two models have been suggested to explain how the recognition of a functional poly(A) signal triggers termination of transcription. One such model implies that co-transcriptional cleavage of the nascent transcript induces transcription termination by allowing the access of XRN2. XRN2 is a 5'-3' exonuclease that operates in degrading the further RNA produced by the still transcribing RNAPII and eventually triggers release of RNAPII from the DNA template (Fong, Brannan et al. 2015, Proudfoot 2016). This has been proven to be the case of β-globin transcription termination which is assisted by a termination sequence element located 800 bp downstream of the poly(A) site (CoTC). Cleavage at the cleavage site and polyadenylation of the nascent transcript is believed to occur later in the nucleoplasm (Nojima, Dienstbier et al. 2013). We measured the duration of fluorescence cycles for the 3' end region of a β-M3 transcripts and obtained values ranging from 15 to 25 seconds. Based on estimated elongation rate of 2-4 kb/min (Singh and Padgett 2009, Larson, Zenklusen et al. 2011, Martin, Rino et al. 2013), RNAPII is expected to take 15-30 seconds to transcribe around 1000 nucleotides from the MS2 stem-loop until the CoTC region. Thus, our data is consistent with release of pre-mRNA β-M3 transcripts occurring at the CoTC. Moreover, it is known that CoTC termination relies on the presence of an upstream poly(A) site, suggesting that the conformational change induced by the assembly of the cleavage and polyadenylation factors is required for CoTC-dependent transcription termination (West, Proudfoot et al. 2008), which is in accordance with our observation that CPSF73 knockdown induces a delay in the release of β-globin transcripts.

When analysing a different construct that does not contain a CoTC element nor a G/C-rich element downstream the poly(A) site (IgM-MPB), we found that it takes a similar time to be released from the transcription site. Thus, the presence of a CoTC element does not seem to influence the kinetics of release of an exogenous reporter genes. Knocking down the cleavage factor CPSF73 by RNAi further supports that termination of IgM minigene depends on a

functional poly(A) signal but does not require cleavage at the poly(A) site, as previously shown for a different gene model (Zhang, Rigo et al. 2015).

The time required to release a transcript from the DNA template at the 3' end therefore becomes a central parameter in our understanding of RNA processing, with implications for both constitutive and alternative polyadenylation that may be disrupted in human genetic disease.

Future perspectives

In summary, our data supports the view that multiple layers of surveillance occur both in the nucleus and in the cytoplasm to minimize potentially toxic effects caused by faulty mRNAs. Although it is not yet possible to predict which splicing mutations will target RNAs for co-transcriptional surveillance or which deep intronic variant will be pathogenic, we expect this work will contribute to open new research venues addressing the impact of splice-sites and deep intronic mutations on mRNA biogenesis in the context of human genetic diseases. Although mRNA analysis is critical for establishing pathogenicity of splice-site, deep intronic and 3' end mutations, biopsy material from affected patient tissues is not always available. This may represent a significant limitation, namely for the study of neurogenetic disorders. Notably, the genes with the longest introns tend to be most highly expressed in the brain (Sibley, Emmett et al. 2015), and non-canonical splicing mechanisms appear enriched these long introns (Roy and Irimia 2008, Pickrell, Pai et al. 2010). Thus, it will be particularly interesting to explore the contribution of deep intronic mutations to human brain disorders and a number of recent possibilities obviate the requirement for brain biopsy. These include using neurons differentiated *in vitro* from either induced pluripotent stem cells (Bellin, Marchetto et al. 2012) or through direct reprogramming (Tsunemoto, Eade et al. 2015). Ultimately, understanding how disease-causing splicing mutations affect mRNA biogenesis may help in the rational design of more effective therapies for these disorders.

Materials and Methods

Cell lines and cell culture

Lymphoblastoid B cell lines

Lymphoblastoid cell lines immortalized by Epstein-Barr virus infection were obtained from the NIGMS Human Genetic Cell Repository collections of the Coriell Institute for Medical Research, USA. Barth syndrome cell lines are GM22129; GM22165; and GM22150. Deafness, autosomal recessive 49 cell lines are GM20190; GM20193; GM20172; GM20189 and Xeroderma Pigmentosum cell line is GM04490. The healthy donor cell line is GM16113. The cell lines are described in detail in Table 1. Cells were cultured in RPMI 1640 medium supplemented with 18% heat-inactivated serum, 2 mM non-essential amino acid solution and 2 mM L-Glutamin at 37°C in 5% CO_2. All cell culture reagents were from Gibco, UK.

HEK 293 cell line

HEK 293 cell line was purchased from Invitrogen Life Technologies and grown as mono- layer in Dulbecco's modified Eagle medium – DMEM (Gibco, UK) supplemented with 10% fetal bovine serum at 37°C in 5% CO_2.

Plasmids

The HBB genomic clone containing a deletion of 593 bp in intron 2 between Rsa I and Ssp I sites was previously described (Antoniou et al 1998). pcDNA5/FRT/TO-RWT4 was constructed by inserting a NcoI-Acc65I (blunted) fragment containing the human HBB gene (from ATG to 1800 bp past the poly(A) site) into the KpnI (blunted) site of pcDNA5/FRT/TO (Invitrogen). The pcDNA5/FRT/TO-EF10 and pcDNA5/FRT/TO-#2MC clones were constructed as the pcDNA5/FRT/TO-RWT4 clone. Additionally, the MS2 fragment was excised by BglII and BamHI digestion, blunted and inserted into either the EcoRI site of HBB exon 3, or the Bam HI site of HBB exon 2. The sequence corresponding to 6 MS2 stem-loops was excised from Pcβwtβ2-6MS2 (Lykke-Andersen et al., 2000) by PCR using the BglII5MS2For and BamHI5MS2Rev pair of primers, followed by BglII and BamHI digestion and insertion into the BglII and BamHI digested pCMV5 vector, generating pCMV5-6MS2. An array of 24 stem-loops was constructed by successive insertions of BglII - BamHI pCMV5-6MS2 fragments into the BglII site of pCMV5-6MS2.

The IgM and IgM-PY minigenes were amplified from pμM (IgM M1-M2) and pPy-AdML-IgM derived constructs described in (Guth et al., 1999 and Guth et al., 2001) using specific primers IgM-Fw and IgM-Rev or IgM-PY-Rev respectively, to introduce Acc65I and BamHI sites and delete a stop codon in exon M2. After Acc65I and BamHI digest the resulting fragments were ligated into the same sites in pECFP-PTS1 generating the vectors pIgM-CFP-PTS1 and pIgM-PY-CFP-PTS. Intron extensions were generated by PCR amplification of fragments from the first intron of mouse RNA Pol II gene using genomic DNA from a murine erythroleukemia (MEL) cell line. The 24 MS2 repeat sequence was ligated into the blunted intronic BbvCI site in pIgM-CFP-PTS1, pIgM-PY-CFP-PTS1, pIgM-600-PY-CFP-PTS1 and into the intronic SwaI site in pIgM-1.7k-PY-CFP-PTS1 respectively. From the pECFP-N1 vector backbone, the final constructs were cut out with HindIII and HpaI and ligated into HindIII and Eco32I sites in pcDNA5/FRT/TO.

The sequence corresponding to 5 λN binding sites (BoxB) was excised from pβglobin.5BoxB (Gehring et al., 2003) by PCR using the BglII5BoxFor and BamHI5BoxRev pair of primers, followed by BglII and BamHI digestion and insertion into the BglII and BamHI sites of pCMV5, generating pCMV5-5BoxB. An array of 25 binding sites was then constructed by successive insertions.

Stable Cell Line Construction

Isogenic, inducible stable cell lines were generated through Flp recombinase-mediated integration by cotransfecting the Flp-InTM T-RExTM-293 (Invitrogen) host cell line harboring a single Flp recombination target site with a plasmid expressing the Flp recombinase (pOG44, Invitrogen) and pcDNA5/FRT/TO-RWT4, pcDNA5/FRT/TO-EF10, pcDNA5/FRT/TO-#2MC or pcDNA5/FRT/TO-IgM constructs at a 9:1 ratio using FuGENE 6 Transfection Reagent (Roche). After transfection, the Flp-In T-REx-293 cells were maintained under selective pressure in the presence of 200 μg/ml hygromycin B (Roche) and 15 μg/ml blasticidin (Invitrogen).

Transient transfections

The Flp-InTM T-RExTM-293 β-WTΔ, β-M2, β-M3, IgM-1.7k-PY-MS2 and IgM-1.7k-PY-MPB cell lines were grown as monolayer in Dulbecco's modified Eagle medium (DMEM) supplemented with 10% fetal bovine serum and 2 mM L-

Glutamin (all cell culture reagents were from Invitrogen). Cells were grown to approximately 70% confluency and transfected 24 hours before imaging with plasmids encoding MS2-GFP, PP7-GFP or λN-GFP fluorescent fusion proteins. Plasmid DNA was transfected using Lipofectamine™ 2000 (Invitrogen) according to the manufacturer's protocol. Plasmid expressing the MS2 coat protein fused to GFP was a gift of E. Bertrand (Boireau et al., 2007 and Fusco et al., 2003) and the λN-GFP and PP7-GFP plasmid was a gift of J. Ellenberg (Daigle and Ellenberg, 2007). Expression of HBB and IgM transgenes was induced with 0.1 µg/ml doxycycline for 2 hours.

RNA interference

Levels of CPSF73 were reduced by SMART pool siRNA (Thermo Scientific). As unspecific siRNA control a sequence targeting the firefly luciferase gene (GL2) was used (Elbashir et al 2001).

Cells were plated in the day before transfection such that they were 40% confluent. The siRNA duplexes were transfected at a final concentration of 20 nM using Lipofectamine RNAiMAX reagent (Life Technologies) according to the manufacturer's protocol and cells were incubated for 72 hours.

Drug treatment

Lymphoblastoid cell lines were treated with 50 µg/ml cycloheximide (C7698, Sigma, USA) for 3h at 37°C.

Total RNA isolation and sub-cellular fractionation

Nuclear and cytoplasmic RNA fractions were isolated as described (37). Briefly, cells were incubated in RSB buffer (10 mM Tris, pH 7.4, 10 mM NaCl, 3 mM$MgCl_2$) for swelling, centrifuged and resuspended in RSBG40 buffer (10 mMTris, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 10% glycerol, 0.2% Nonidet P-40 (lymphoblastoid cells) or 0.5% Nonidet P-40 (HEK 293 cells), 0.5 mM dithiothreitol and 40 U/ml RNase) for lyses of the cell membrane. The fractionation of the nuclei into chromatin-associated and nucleoplasmic RNA was adapted from (13-15). The nuclear pellet was gently resuspended in a prechilled glycerol buffer (20 mM Tris pH 7.9, 75 mM NaCl, 0.5 mM EDTA, 0.85 mM DTT, 0.125 mM PMSF and 50% glycerol) and an equal volume of cold nuclei lysis buffer (10 mM HEPES pH7.6, 300 mM NaCl, 0.2 mM EDTA, 1

mM DTT, 7.5 mM $MgCl_2$, 1 M Urea and 1% NP-40) was added. The tube was gently vortexed for 2 × 2 s and incubated for 10 min on ice. Chromatin was pelleted and incubated in 10 mM Tris pH 7.5, 500 mM NaCl, 10 mM $MgCl_2$, 100 U/µl DNase I, 100 U/µl RNase OUT. RNA was extracted from each fraction and from the whole cell using PureZOL RNA isolation reagent (Bio-Rad, USA) and treated with RNase-free DNaseI (Roche).

4sU labelling of nascent transcript

4-thiouridine (4sU, Sigma) was added to the cells and made up to a final concentration of 500 µM. Lymphoblastoid B cells were incubated with 4sU for 10 minutes and HEK 293 cells for 2 or 60 minutes, total RNA was immediately extracted using PureZOL RNA isolation reagent (Bio-Rad, USA).

Purification of 4sU-labelled RNA

4sU-labelled RNA extracted from lymphoblastoid B cells was purified using HPDP-biotin as described in Friedel and Dolken (Friedel and Dolken 2009).

Regarding HEK 293 cells, 4sU-labelled RNA was purified using HPDP-biotin or MTS-biotin, following protocols adapted from Friedel and Dolken (Friedel and Dolken 2009) and from Duffy et al. (Duffy, Rutenberg-Schoenberg et al. 2015). Shortly, the thiol-labelled RNA was biotinylated using EZ-Link HPDP-biotin (Pierce, USA) or MTS-biotin (Biotium, USA) and separated from untagged species using µMACS streptavidin coated magnetic beads and columns (Miltenyi, Germany). Biotinylation reactions were carried out in a total volume of 1 ml (HPDP-biotin) or 250 µl (MTS-biotin) and incubated at room temperature for 90 minutes. All reaction contained 100 µg total RNA, 10 mM HEPES (pH7.5), 1 mM EDTA, and 10 mg MTS-biotin or 200 mg HPDP-biotin both dissolved in DMF. Two rounds of purification of biotinylated RNA were performed (phenol:chloroform and chloroform:isoamylalcohol). Biotinylated RNA was then precipitated by adding 5M NaCl and isopropanol. Labelled RNA was separated from the pre-existing RNA using 200 µl µMACS streptavidin beads and µColumns. 4sU-labelled RNA was eluted from the columns with 100 mM DTT and precipitated by adding ethanol.

RT-PCR

DNAse-treated RNA was used as template for cDNA synthesis using random primers from the Transcriptor High Fidelity cDNA Synthesis Kit (Roche) according to the manufacturer's instructions. cDNA products were amplified by NZYLong DNA polymerase (NZYTech, Portugal) and were separated by agarose gel electrophoresis, detected by GelRed (Biotium, Inc., USA), and imaged on the Chemido XRS+ (Bio-Rad, USA). Analysis of PCR product abundance was carried out on signal intensities from each band of non-saturated gel images using ImageJ software. Gene-specific primers are presented in Table M1 for Lymphoblastoid B cells. Additionally, gene-specific primers are presented in Table M2 and M3 for HEK 293 cells.

Quantitative real-time PCR

DNAse-treated RNA was used as template for cDNA synthesis using random primers from the Transcriptor High Fidelity cDNA Synthesis Kit (Roche) according to the manufacturer's instructions. PCR reactions were performed in the ViiA™ 7 Real-Time PCR System (Applied Biosystems, USA), using iTaq Universal SYBR Green Supermix (Bio-Rad, USA). Gene-specific primers are presented in Table M1 for Lymphoblastoid B cells and in Table M2 for HEK2093 cells. Each sample was run in duplicate. The $2^{-\Delta Ct}$ method (Schmittgen and Livak 2008) was used to measure the relative changes in transcript levels.

Immunoblotting

Proteins were isolated from chromatin, nucleoplasmic and cytoplasmic fraction of lymphoblastoid cells as previously described (de Almeida, Garcia-Sacristan et al. 2010). Equal amounts of protein extracts were resolved by SDS-polyacrylamide gel electrophoresis and transferred to a nitrocellulose membrane that was subsequently incubated with the following primary antibodies: anti-lamin A/C (H-110, Santa Cruz Biotechnology, Inc); anti β-actin (Sigma); anti-U2B″ (clone 4G3, PROGEN Biotechnik GmbH); and anti-histone H3 (Abcam).

Whole-cell lysates of HEK 293 cells were prepared and incubated with CPSF73 (Bethyl laboratories) and tubulin (Sigma) primary antibodies followed by

incubation with the appropriate secondary antibodies (BioRad) and by detection using enhanced luminescence (Amersham).

Microarray data analysis

Data deposited in GEO database with the reference GSE34204 (Duan, Shi et al. 2013), was used for analysis. Microarrays were processed by using the AltAnalyze software version 2.0.8 (Emig, Salomonis et al. 2010). Briefly, raw CEL data files from the deposited microarrays were normalized by the RMA algorithm. Probesets with detection above background (DABG) p-values above 0.5 or non-logarithmic expression below 1.0 were removed from the analysis. Gene expression levels were determined using only constitutive probesets, using the gene annotation present in AltAnalyze derived from Ensembl (Flicek, Aken et al. 2008) and USCS (Meyer, Zweig et al. 2013) databases.

Spinning-Disk Confocal Live-Cell Imaging

Cells were plated on 25 mm diameter glass coverslips coated with 0.01% Poly-L-Lysine. Before imaging, the medium was changed to α-MEM without phenol red (Invitrogen) supplemented with 20 mM HEPES, pH 7.4 and 10% FBS. Each coverslip was mounted into a perfusion chamber and placed in a heated sample holder (20/20 Technology, Inc.; Wilmington, NC) mounted on the stage of a Marianas™ imaging system (Intelligent Imaging Innovations, Denver, CO) based on an Axio Observer inverted microscope (Carl Zeiss, Inc.; Thornwood, NY) equipped with a spinning-disk confocal head (Yokogawa Electric, Tokyo, Japan). The microscope stage and objective lenses were maintained inside an environmental chamber (Okolab) set at 37°C with 100% humidity. The axial position of the sample was controlled with a piezo-driven stage (Applied Scientific Instrumentation, Eugene, OR). Samples were illuminated using 70 % mW solid state lasers (λ = 488 nm for GFP, Coherent, Inc.; Santa Clara, CA) coupled to an acoustic-optical tunable filter (AOTF), with Gain 1, Speed 1 and Intensification 300. Images were acquired using 100x (Plan-Apo, 1.4 NA) oil immersion objectives (Carl Zeiss, Inc.) under control of Slidebook 6.0 software (Intelligent Imaging Innovations, Denver, CO). Digital images (16-bit) were obtained using a cooled CCD camera (QuantEM, 512SC, Photometrics, Tucson, AZ) with acquisition time of 30 ms (for transcription sites in the cell nucleus), average timelapse interval of 5s, z-stacks of 8 steps with 0.27 µm size each.

Image Analysis

Image analysis was performed as described in (Rino, de Jesus et al. 2017). Briefly, the following steps were performed in the analysis of each single transcription site in a time lapse sequence. (1) The XY position of the transcription site in all time frame of the sequence was determined as a local of maximum intensity projection image. (2) The total fluorescence intensity of the transcription site was calculated for each time point by performing a 2D Gaussian fit at the Z plane corresponding to the highest intensity value. To determine the average TFI for a single transcript, we used the same software to analyse images of RNA molecules diffusing in the nucleoplasm of cells acquired with an exposure time of 30 ms. Nucleoplasmic-released mRNA molecules are no longer immobilized at the transcription site but diffuse throughout the nucleoplasm, we performed quantification in single optical planes. We used MATLAB to plot histograms of TFI and perform Gaussian fitting on the observed distributions to determine the TFI values which are within one standard deviation (68%) and two standard deviations (95%) away from the mean. Since the EMCCD camera output is linearly dependent on the exposure time, the average TFI corresponding to a single intron for a different exposure time can be obtained by multiplying the TFI at 30 ms by the ratio T/30, where T is the new exposure time in ms. The number of fluorescent introns at a given transcription site is then estimated by dividing the TFI of the transcription site by the average TFI of a single intron for the same exposure time.

Gene	Primer Name	Sequence
	RT-qPCR primers	
GAPDH	GAPDH For	GAAGGTGGAGGTCGGAGTC
	GAPDH Rev	GAAGATGGTGATGGGATTTC
TAZ	TAZ ex.4 For	AGACATCTGCTTCACCAAGGAGCTA
	TAZ ex.4 Rev	TCGGCACACAGGCACACACT
	TAZ ex. 11 For	TGCGGAAAGCCCTGACGGA
	TAZ ex. 11 Rev	GGCTGGAGGTGGTTGTGGAGC
	TAZ int.10 For	GCCTCCACCCTCTCCATCCCG
	TAZ int.10 Rev	TGCACCCCTCGGGAAGCTTGG
MARVELD2	MARVELD2 ex.2 For	CTCCAGCAAGACCAAACCAC
	MARVELD2 ex.2 Rev	CAGCCTCTTTCCGGGAACTA
	MARVELD2 ex.4 For	GGTGACAGACAAAGAGACTCAG
	MARVELD2 ex.4 Rev	ACATAGTCGGGCATCACGAT
	MARVELD2 int.3 For	AGGTGATCTGGCTTCTGTCC
	MARVELD2 int.3 Rev	TGGATTAGGTGTGGAGGCTG
XPC	XPC ex. 1 For	GGCCGGCGTTCTAGCGCAT
	XPC ex. 1 Rev	CACGCCGGGCCTTGCTCTTG
	XPC ex. 10 For	GGCTAAACACATGGACCAGC
	XPC ex. 10 Rev	GTAGACCGCTTCTCCACGAC
	XPC ex.11_int.11	AGGCTTGGAGAAGTACCCTACAAG
	XPC ex.11_int.11	TGAATCCTGCTCAAGCCGGGAAA
	RT-PCR primers	
GAPDH	GAPDH ex.3 For	TCACCAGGGCTGCTTTTAAC
	GAPDH ex.3 Rev	CATGTAGTTGAGGTCAATGAAGG
	GAPDH ex.5 Rev	TGAAGACGCCAGTGGAC
	GAPDH int.2 For	GGGAAGGAAATGAATGGGCAG
	GAPDH int.2 Rev	GGACCTCCATAAACCCACTT
XIST	Xist For	GTCAGGAGAAAGAAGTGGAGGG
	Xist Rev	ACAGAGGAATGGAGGGAGGTT

Table M1

Gene	Primer name	Total, UnSpliced, SPliced	Sequence
COL4A6	int.3_ex.4 For	US	AATTGCACTCCTTCTGTCCAC
	int.3_ex.4 Rev	US	ACCCCTTTCTCCTTTCAATCC
	ex.3_ex.4 For	SP	GGAGCTGTCAGTGTTTTCCTG
	ex.3_ex.4 Rev	SP	TTTCAATCCCGATAAACCAGTAG
HPRT1	ex.1_int.1 For	US	CTTCCTCCTCCTGAGCAGTC
	ex.1_int.1 Rev	US	CGAACCCGGGAAACTGG
	ex.1 For	SP	GCGAACCTCTCGGCTTTC
	int.1_ex.2 Rev	SP	ACCCTTTCCAAATCCTCAGC
CTNNBL1	ex.1_int.1 For	US	GTGAGTGCAGGGAAGTGGAG
	ex.1_int.1 Rev	US	GGTGAGATGAAAGGGCTCTG
	ex.1 For	SP	CCGCACTTTACGGCAGTG
	int1_ex.2 Rev	SP	TCATTTCTTCTTCCCGATAGC
GAPDH	ex.1-int.1 For	US	CTCCTGTTCGACAGTCAGC
	ex.1-int.1 Rev	US	TTCAGGCCGTCCCTAGC
	ex.2-ex.4 For	SP	GAAGGTGGAGGTCGGAGTC
	ex.2-ex.4 Rev	SP	GAAGATGGTGATGGGATTTC
U6	For	T	GCTTCGGCAGCACATATACTA
	Rev	T	AAATATGGAACGCTTCACGA

Table M2

Gene	Primer Name	Sequence
	RT-qPCR primers	
HBB	Ex.1 For	ACGTGGATGAAGTTGGTGGT
	Ex. 2 Rev	CACCTTTGCCACACTGAGTG
	Int. 1 For	TCTTGGGTTTCTGATAGGCAC
	Ex. 2 For	CTGCTGGTGGTCTACCCTTG
	Ex. 3 Rev	CAACGTGCTGGTCTGTGTG
	Int. 2 For	AGTTCATGTCATAGGAAGGGGAGAAG
IgM mini gene	Ex. 1 For	GAATTCTGCAGTCGACGGTAC
	Int. 1 For	AGAGGTGTTTGAGGACACAGG
	Ex. 2 For	CGTCTATATCACCGCCGACA
	Ex. 2 Rev	TGGGTGCTCAGGTAGTGGTT
	UC For	AGCCTCGACTGTGCCTTCTA
	UC Rev	CCCAGAATAGAATGACACCTACTCA
	UT For	CTGGGGCTCTAGGGGGTAT
	UT Rev	TTTAGAGCTTGACGGGGAAA
	RT-PCR primers	
HBB	Ex.1 For	TACCATGGTGCACCTGACTC
	Ex. 1 Rev	AGTTGGTGGTGAGGCCCT
	Ex. 3 For	AAGCTCGCTTTCTTGCTGTC
	Ex. 3 Rev	GCCTTGAGCATCTGGATTCTGCC

Table M3

References

Aitken, S., R. D. Alexander and J. D. Beggs (2011). "Modelling reveals kinetic advantages of co-transcriptional splicing." PLoS Comput Biol 7(10): e1002215.

Akman, H. O., O. Kakhlon, J. Coku, L. Peverelli, H. Rosenmann, L. Rozenstein-Tsalkovich, J. Turnbull, V. Meiner, L. Chama, I. Lerer, S. Shpitzen, E. Leitersdorf, C. Paradas, M. Wallace, R. Schiffmann, S. DiMauro, A. Lossos and B. A. Minassian (2015). "Deep intronic GBE1 mutation in manifesting heterozygous patients with adult polyglucosan body disease." JAMA Neurol 72(4): 441-445.

Alexander, R. D., S. A. Innocente, J. D. Barrass and J. D. Beggs (2010). "Splicing-dependent RNA polymerase pausing in yeast." Mol Cell 40(4): 582-593.

Allen, B. L. and D. J. Taatjes (2015). "The Mediator complex: a central integrator of transcription." Nat Rev Mol Cell Biol 16(3): 155-166.

Almada, A. E., X. Wu, A. J. Kriz, C. B. Burge and P. A. Sharp (2013). "Promoter directionality is controlled by U1 snRNP and polyadenylation signals." Nature 499(7458): 360-363.

Ameur, A., A. Zaghlool, J. Halvardson, A. Wetterbom, U. Gyllensten, L. Cavelier and L. Feuk (2011). "Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain." Nat Struct Mol Biol 18(12): 1435-1440.

Amorim, M. J., C. Cotobal, C. Duncan and J. Mata (2010). "Global coordination of transcriptional control and mRNA decay during cellular differentiation." Mol Syst Biol 6: 380.

Antonellis, A., M. Y. Dennis, G. Burzynski, J. Huynh, V. Maduro, C. J. Hodonsky, M. Khajavi, K. Szigeti, S. Mukkamala, S. L. Bessling, W. J. Pavan, A. S. McCallion, J. R. Lupski, E. D. Green and N. C. S. Program (2010). "A rare myelin protein zero (MPZ) variant alters enhancer activity in vitro and in vivo." PLoS One 5(12): e14346.

Aslanzadeh, V., Y. Huang, G. Sanguinetti and J. D. Beggs (2018). "Transcription rate strongly affects splicing fidelity and cotranscriptionality in budding yeast." Genome Res 28(2): 203-213.

Banerjee, A., L. H. Apponi, G. K. Pavlath and A. H. Corbett (2013). "PABPN1: molecular function and muscle disease." FEBS J 280(17): 4230-4250.

Barash, Y., J. A. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, B. J. Blencowe and B. J. Frey (2010). "Deciphering the splicing code." Nature 465(7294): 53-59.

Barrass, J. D., J. E. Reid, Y. Huang, R. D. Hector, G. Sanguinetti, J. D. Beggs and S. Granneman (2015). "Transcriptome-wide RNA processing kinetics revealed using extremely short 4tU labeling." Genome Biol 16: 282.

Bayne, E. H., M. Portoso, A. Kagansky, I. C. Kos-Braun, T. Urano, K. Ekwall, F. Alves, J. Rappsilber and R. C. Allshire (2008). "Splicing factors facilitate RNAi-directed silencing in fission yeast." Science 322(5901): 602-606.

Becker, P. W., N. Sacilotto, S. Nornes, A. Neal, M. O. Thomas, K. Liu, C. Preece, I. Ratnayaka, B. Davies, G. Bou-Gharios and S. De Val (2016). "An Intronic Flk1 Enhancer Directs Arterial-Specific Expression via RBPJ-Mediated Venous Repression." Arterioscler Thromb Vasc Biol **36**(6): 1209-1219.

Belair, C., S. Sim and S. L. Wolin (2017). "Noncoding RNA Surveillance: The Ends Justify the Means." Chem Rev.

Bellin, M., M. C. Marchetto, F. H. Gage and C. L. Mummery (2012). "Induced pluripotent stem cells: the new patient?" Nat Rev Mol Cell Biol **13**(11): 713-726.

Beltran, M., I. Puig, C. Pena, J. M. Garcia, A. B. Alvarez, R. Pena, F. Bonilla and A. G. de Herreros (2008). "A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition." Genes Dev **22**(6): 756-769.

Bentley, D. L. (2014). "Coupling mRNA processing with transcription in time and space." Nat Rev Genet **15**(3): 163-175.

Berezikov, E., W. J. Chung, J. Willis, E. Cuppen and E. C. Lai (2007). "Mammalian mirtron genes." Mol Cell **28**(2): 328-336.

Berglund, J. A., K. Chua, N. Abovich, R. Reed and M. Rosbash (1997). "The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC." Cell **89**(5): 781-787.

Bernecky, C., F. Herzog, W. Baumeister, J. M. Plitzko and P. Cramer (2016). "Structure of transcribing mammalian RNA polymerase II." Nature **529**(7587): 551-554.

Beroud, C., A. Carrie, C. Beldjord, N. Deburgrave, S. Llense, N. Carelle, C. Peccate, J. M. Cuisset, F. Pandit, F. Carre-Pigeon, M. Mayer, R. Bellance, D. Recan, J. Chelly, J. C. Kaplan and F. Leturcq (2004). "Dystrophinopathy caused by mid-intronic substitutions activating cryptic exons in the DMD gene." Neuromuscul Disord **14**(1): 10-18.

Beyer, A. L. and Y. N. Osheim (1988). "Splice site selection, rate of splicing, and alternative splicing on nascent transcripts." Genes Dev **2**(6): 754-765.

Bhatt, D. M., A. Pandya-Jones, A. J. Tong, I. Barozzi, M. M. Lissner, G. Natoli, D. L. Black and S. T. Smale (2012). "Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions." Cell **150**(2): 279-290.

Bieberstein, N. I., F. Carrillo Oesterreich, K. Straube and K. M. Neugebauer (2012). "First exon length controls active chromatin signatures and transcription." Cell Rep **2**(1): 62-68.

Bione, S., P. D'Adamo, E. Maestrini, A. K. Gedeon, P. A. Bolhuis and D. Toniolo (1996). "A novel X-linked gene, G4.5. is responsible for Barth syndrome." Nat Genet **12**(4): 385-389.

Bjork, P. and L. Wieslander (2017). "Integration of mRNP formation and export." Cell Mol Life Sci **74**(16): 2875-2897.

Black, D. L. (2003). "Mechanisms of alternative pre-messenger RNA splicing." Annu Rev Biochem 72: 291-336.

Boireau, S., P. Maiuri, E. Basyuk, M. de la Mata, A. Knezevich, B. Pradet-Balade, V. Backer, A. Kornblihtt, A. Marcello and E. Bertrand (2007). "The transcriptional cycle of HIV-1 in real-time and live cells." J Cell Biol 179(2): 291-304.

Borbolis, F. and P. Syntichaki (2015). "Cytoplasmic mRNA turnover and ageing." Mech Ageing Dev 152: 32-42.

Bortolin, M. L. and T. Kiss (1998). "Human U19 intron-encoded snoRNA is processed from a long primary transcript that possesses little potential for protein coding." RNA 4(4): 445-454.

Boutz, P. L., A. Bhutkar and P. A. Sharp (2015). "Detained introns are a novel, widespread class of post-transcriptionally spliced introns." Genes Dev 29(1): 63-80.

Bresson, S. M. and N. K. Conrad (2013). "The human nuclear poly(a)-binding protein promotes RNA hyperadenylation and decay." PLoS Genet 9(10): e1003893.

Brinster, R. L., J. M. Allen, R. R. Behringer, R. E. Gelinas and R. D. Palmiter (1988). "Introns increase transcriptional efficiency in transgenic mice." Proc Natl Acad Sci U S A 85(3): 836-840.

Brody, Y., N. Neufeld, N. Bieberstein, S. Z. Causse, E. M. Bohnlein, K. M. Neugebauer, X. Darzacq and Y. Shav-Tal (2011). "The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing." PLoS Biol 9(1): e1000573.

Brugiolo, M., L. Herzel and K. M. Neugebauer (2013). "Counting on co-transcriptional splicing." F1000Prime Rep 5: 9.

Buratti, E., M. Chivers, G. Hwang and I. Vorechovsky (2011). "DBASS3 and DBASS5: databases of aberrant 3'- and 5'-splice sites." Nucleic Acids Res 39(Database issue): D86-91.

Buratti, E., M. Chivers, J. Kralovicova, M. Romano, M. Baralle, A. R. Krainer and I. Vorechovsky (2007). "Aberrant 5' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization." Nucleic Acids Res 35(13): 4250-4263.

Buratti, E., A. Dhir, M. A. Lewandowska and F. E. Baralle (2007). "RNA structure is a key regulatory element in pathological ATM and CFTR pseudoexon inclusion events." Nucleic Acids Res 35(13): 4369-4383.

Burger, K., B. Muhl, M. Kellner, M. Rohrmoser, A. Gruber-Eber, L. Windhager, C. C. Friedel, L. Dolken and D. Eick (2013). "4-thiouridine inhibits rRNA synthesis and causes a nucleolar stress response." RNA Biol 10(10).

Burnette, J. M., E. Miyamoto-Sato, M. A. Schaub, J. Conklin and A. J. Lopez (2005). "Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements." Genetics 170(2): 661-674.

Busslinger, M., N. Moschonas and R. A. Flavell (1981). "Beta + thalassemia: aberrant splicing results from a single point mutation in an intron." Cell 27(2 Pt 1): 289-298.

Bussow, K. (2015). "Stable mammalian producer cell lines for structural biology." Curr Opin Struct Biol 32: 81-90.

Calado, A., F. M. Tome, B. Brais, G. A. Rouleau, U. Kuhn, E. Wahle and M. Carmo-Fonseca (2000). "Nuclear inclusions in oculopharyngeal muscular dystrophy consist of poly(A) binding protein 2 aggregates which sequester poly(A) RNA." Hum Mol Genet 9(15): 2321-2328.

Caminsky, N., E. J. Mucaki and P. K. Rogan (2014). "Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis." F1000Res 3: 282.

Carrillo Oesterreich, F., S. Preibisch and K. M. Neugebauer (2010). "Global analysis of nascent RNA reveals transcriptional pausing in terminal exons." Mol Cell 40(4): 571-581.

Carvalho, T., S. Martins, J. Rino, S. Marinho and M. Carmo-Fonseca (2017). "Pharmacological inhibition of the spliceosome subunit SF3b triggers exon junction complex-independent nonsense-mediated decay." J Cell Sci 130(9): 1519-1531.

Chabot, B. and L. Shkreta (2016). "Defective control of pre-messenger RNA splicing in human disease." J Cell Biol 212(1): 13-27.

Chang, H., J. Lim, M. Ha and V. N. Kim (2014). "TAIL-seq: genome-wide determination of poly(A) tail length and 3' end modifications." Mol Cell 53(6): 1044-1052.

Chen, H., Y. Huo, S. Patel, X. Zhu, T. Swift-Scanlan, R. H. Reeves, R. DePaulo, Jr., C. A. Ross and M. G. McInnis (2000). "Gene identification using exon amplification on human chromosome 18q21: implications for bipolar disorder." Mol Psychiatry 5(5): 502-509.

Chen, L. L. (2016). "The biogenesis and emerging roles of circular RNAs." Nat Rev Mol Cell Biol 17(4): 205-211.

Chen, M. and J. L. Manley (2009). "Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches." Nat Rev Mol Cell Biol 10(11): 741-754.

Chorev, M. and L. Carmel (2013). "Computational identification of functional introns: high positional conservation of introns that harbor RNA genes." Nucleic Acids Res 41(11): 5604-5613.

Chow, L. T., R. E. Gelinas, T. R. Broker and R. J. Roberts (1977). "An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA." Cell 12(1): 1-8.

Cleary, M. D., C. D. Meiering, E. Jan, R. Guymon and J. C. Boothroyd (2005). "Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay." Nat Biotechnol **23**(2): 232-237.

Close, P., P. East, A. B. Dirac-Svejstrup, H. Hartmann, M. Heron, S. Maslen, A. Chariot, J. Soding, M. Skehel and J. Q. Svejstrup (2012). "DBIRD complex integrates alternative mRNA splicing with RNA polymerase II transcript elongation." Nature **484**(7394): 386-389.

Consortium, E. P. (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.

Core, L. J., J. J. Waterfall and J. T. Lis (2008). "Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters." Science **322**(5909): 1845-1848.

Corvelo, A. and E. Eyras (2008). "Exon creation and establishment in human genes." Genome Biol **9**(9): R141.

Coulon, A., C. C. Chow, R. H. Singer and D. R. Larson (2013). "Eukaryotic transcriptional dynamics: from single molecules to cell populations." Nat Rev Genet **14**(8): 572-584.

Coulon, A., M. L. Ferguson, V. de Turris, M. Palangat, C. C. Chow and D. R. Larson (2014). "Kinetic competition during the transcription cycle results in stochastic RNA processing." Elife **3**.

Cramer, P. (2016). "Structure determination of transient transcription complexes." Biochem Soc Trans **44**(4): 1177-1182.

Cunha, K. S., N. S. Oliveira, A. K. Fausto, C. C. de Souza, A. Gros, T. Bandres, Y. Idrissi, J. P. Merlio, R. S. de Moura Neto, R. Silva, M. Geller and D. Cappellen (2016). "Hybridization Capture-Based Next-Generation Sequencing to Evaluate Coding Sequence and Deep Intronic Mutations in the NF1 Gene." Genes (Basel) **7**(12).

Curado, J., C. Iannone, H. Tilgner, J. Valcarcel and R. Guigo (2015). "Promoter-like epigenetic signatures in exons displaying cell type-specific splicing." Genome Biol **16**: 236.

Curinha, A., S. Oliveira Braz, I. Pereira-Castro, A. Cruz and A. Moreira (2014). "Implications of polyadenylation in health and disease." Nucleus **5**(6): 508-519.

Cusack, B. P., P. F. Arndt, L. Duret and H. Roest Crollius (2011). "Preventing dangerous nonsense: selection for robustness to transcriptional error in human genes." PLoS Genet **7**(10): e1002276.

Custodio, N. and M. Carmo-Fonseca (2016). "Co-transcriptional splicing and the CTD code." Crit Rev Biochem Mol Biol **51**(5): 395-411.

Custodio, N., M. Carmo-Fonseca, F. Geraghty, H. S. Pereira, F. Grosveld and M. Antoniou (1999). "Inefficient processing impairs release of RNA from the site of transcription." EMBO J **18**(10): 2855-2866.

Damgaard, C. K., S. Kahns, S. Lykke-Andersen, A. L. Nielsen, T. H. Jensen and J. Kjems (2008). "A 5' splice site enhances the recruitment of basal transcription initiation factors in vivo." Mol Cell **29**(2): 271-278.

Danckwardt, S., M. W. Hentze and A. E. Kulozik (2008). "3' end mRNA processing: molecular mechanisms and implications for health and disease." EMBO J **27**(3): 482-498.

Danko, C. G., N. Hah, X. Luo, A. L. Martins, L. Core, J. T. Lis, A. Siepel and W. L. Kraus (2013). "Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells." Mol Cell **50**(2): 212-222.

Dar, R. D., B. S. Razooky, A. Singh, T. V. Trimeloni, J. M. McCollum, C. D. Cox, M. L. Simpson and L. S. Weinberger (2012). "Transcriptional burst frequency and burst size are equally modulated across the human genome." Proc Natl Acad Sci U S A **109**(43): 17454-17459.

Darzacq, X., Y. Shav-Tal, V. de Turris, Y. Brody, S. M. Shenoy, R. D. Phair and R. H. Singer (2007). "In vivo dynamics of RNA polymerase II transcription." Nat Struct Mol Biol **14**(9): 796-806.

David, C. J., A. R. Boyne, S. R. Millhouse and J. L. Manley (2011). "The RNA polymerase II C-terminal domain promotes splicing activation through recruitment of a U2AF65-Prp19 complex." Genes Dev **25**(9): 972-983.

Davidson, L., A. Kerr and S. West (2012). "Co-transcriptional degradation of aberrant pre-mRNA by Xrn2." EMBO J **31**(11): 2566-2578.

de Almeida, S. F., A. Garcia-Sacristan, N. Custodio and M. Carmo-Fonseca (2010). "A link between nuclear RNA surveillance, the human exosome and RNA polymerase II transcriptional termination." Nucleic Acids Res **38**(22): 8015-8026.

de Almeida, S. F., A. R. Grosso, F. Koch, R. Fenouil, S. Carvalho, J. Andrade, H. Levezinho, M. Gut, D. Eick, I. Gut, J. C. Andrau, P. Ferrier and M. Carmo-Fonseca (2011). "Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36." Nat Struct Mol Biol **18**(9): 977-983.

De Conti, L., M. Baralle and E. Buratti (2013). "Exon and intron definition in pre-mRNA splicing." Wiley Interdiscip Rev RNA **4**(1): 49-60.

de la Mata, M., C. R. Alonso, S. Kadener, J. P. Fededa, M. Blaustein, F. Pelisch, P. Cramer, D. Bentley and A. R. Kornblihtt (2003). "A slow RNA polymerase II affects alternative splicing in vivo." Mol Cell **12**(2): 525-532.

Dhir, A. and E. Buratti (2010). "Alternative splicing: role of pseudoexons in human disease and potential therapeutic strategies." FEBS J 277(4): 841-855.

Dobkin, C., R. G. Pergolizzi, P. Bahre and A. Bank (1983). "Abnormal splice in a mutant human beta-globin gene not at the site of a mutation." Proc Natl Acad Sci U S A 80(5): 1184-1188.

Dolken, L., Z. Ruzsics, B. Radle, C. C. Friedel, R. Zimmer, J. Mages, R. Hoffmann, P. Dickinson, T. Forster, P. Ghazal and U. H. Koszinowski (2008). "High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay." RNA 14(9): 1959-1972.

Dong, R., X. K. Ma, L. L. Chen and L. Yang (2016). "Increased complexity of circRNA expression during species evolution." RNA Biol: 1-11.

Dreyfuss, G., V. N. Kim and N. Kataoka (2002). "Messenger-RNA-binding proteins and the messages they carry." Nat Rev Mol Cell Biol 3(3): 195-205.

Duan, J., J. Shi, X. Ge, L. Dolken, W. Moy, D. He, S. Shi, A. R. Sanders, J. Ross and P. V. Gejman (2013). "Genome-wide survey of interindividual differences of RNA stability in human lymphoblastoid cell lines." Sci Rep 3: 1318.

Duff, M. O., S. Olson, X. Wei, S. C. Garrett, A. Osman, M. Bolisetty, A. Plocik, S. E. Celniker and B. R. Graveley (2015). "Genome-wide identification of zero nucleotide recursive splicing in Drosophila." Nature 521(7552): 376-379.

Duffy, E. E., M. Rutenberg-Schoenberg, C. D. Stark, R. R. Kitchen, M. B. Gerstein and M. D. Simon (2015). "Tracking Distinct RNA Populations Using Efficient and Reversible Covalent Chemistry." Mol Cell 59(5): 858-866.

Dumesic, P. A., P. Natarajan, C. Chen, I. A. Drinnenberg, B. J. Schiller, J. Thompson, J. J. Moresco, J. R. Yates, 3rd, D. P. Bartel and H. D. Madhani (2013). "Stalled spliceosomes are a signal for RNAi-mediated genome defense." Cell 152(5): 957-968.

Dye, M. J., N. Gromak and N. J. Proudfoot (2006). "Exon tethering in transcription by RNA polymerase II." Mol Cell 21(6): 849-859.

Dye, M. J. and N. J. Proudfoot (2001). "Multiple transcript cleavage precedes polymerase release in termination by RNA polymerase II." Cell 105(5): 669-681.

Eberle, A. B. and N. Visa (2014). "Quality control of mRNP biogenesis: networking at the transcription site." Semin Cell Dev Biol 32: 37-46.

Edery, P., C. Marcaillou, M. Sahbatou, A. Labalme, J. Chastang, R. Touraine, E. Tubacher, F. Senni, M. B. Bober, S. Nampoothiri, P. S. Jouk, E. Steichen, S. Berland, A. Toutain, C. A. Wise, D. Sanlaville, F. Rousseau, F. Clerget-Darpoux and A. L. Leutenegger (2011). "Association of TALS developmental disorder with defect in minor splicing component U4atac snRNA." Science 332(6026): 240-243.

Ehrensberger, A. H., G. P. Kelly and J. Q. Svejstrup (2013). "Mechanistic interpretation of promoter-proximal peaks and RNAPII density maps." Cell 154(4): 713-715.

Emig, D., N. Salomonis, J. Baumbach, T. Lengauer, B. R. Conklin and M. Albrecht (2010). "AltAnalyze and DomainGraph: analyzing and visualizing exon expression data." Nucleic Acids Res 38(Web Server issue): W755-762.

Emili, A., M. Shales, S. McCracken, W. Xie, P. W. Tucker, R. Kobayashi, B. J. Blencowe and C. J. Ingles (2002). "Splicing and transcription-associated proteins PSF and p54nrb/nonO bind to the RNA polymerase II CTD." RNA 8(9): 1102-1111.

Engreitz, J. M., A. Pandya-Jones, P. McDonel, A. Shishkin, K. Sirokman, C. Surka, S. Kadri, J. Xing, A. Goren, E. S. Lander, K. Plath and M. Guttman (2013). "The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome." Science 341(6147): 1237973.

Flicek, P., B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, T. Eyre, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, R. Holland, K. L. Howe, K. Howe, N. Johnson, A. Jenkinson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. J. Vilella, J. Vogel, S. White, M. Wood, E. Birney, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, T. J. Hubbard, A. Kasprzyk, G. Proctor, J. Smith, A. Ureta-Vidal and S. Searle (2008). "Ensembl 2008." Nucleic Acids Res 36(Database issue): D707-714.

Flynt, A. S., J. C. Greimann, W. J. Chung, C. D. Lima and E. C. Lai (2010). "MicroRNA biogenesis via splicing and exosome-mediated trimming in Drosophila." Mol Cell 38(6): 900-907.

Fong, N., K. Brannan, B. Erickson, H. Kim, M. A. Cortazar, R. M. Sheridan, T. Nguyen, S. Karp and D. L. Bentley (2015). "Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition." Mol Cell 60(2): 256-267.

Fong, N., H. Kim, Y. Zhou, X. Ji, J. Qiu, T. Saldi, K. Diener, K. Jones, X. D. Fu and D. L. Bentley (2014). "Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate." Genes Dev 28(23): 2663-2676.

Friedel, C. C. and L. Dolken (2009). "Metabolic tagging and purification of nascent RNA: implications for transcriptomics." Mol Biosyst 5(11): 1271-1278.

Fuchs, G., Y. Voichek, S. Benjamin, S. Gilad, I. Amit and M. Oren (2014). "4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells." Genome Biol 15(5): R69.

Fuchs, G., Y. Voichek, M. Rabani, S. Benjamin, S. Gilad, I. Amit and M. Oren (2015). "Simultaneous measurement of genome-wide transcription elongation speeds and rates

of RNA polymerase II transition into active elongation with 4sUDRB-seq." Nat Protoc **10**(4): 605-618.

Fusco, D., N. Accornero, B. Lavoie, S. M. Shenoy, J. M. Blanchard, R. H. Singer and E. Bertrand (2003). "Single mRNA molecules demonstrate probabilistic movement in living mammalian cells." Curr Biol **13**(2): 161-167.

Gaffney, D. J. and P. D. Keightley (2004). "Unexpected conserved non-coding DNA blocks in mammals." Trends Genet **20**(8): 332-337.

Gazzoli, I., I. Pulyakhina, N. E. Verwey, Y. Ariyurek, J. F. Laros, P. A. t Hoen and A. Aartsma-Rus (2016). "Non-sequential and multi-step splicing of the dystrophin transcript." RNA Biol **13**(3): 290-305.

Ghosh, S. and A. Jacobson (2010). "RNA decay modulates gene expression and controls its fidelity." Wiley Interdiscip Rev RNA **1**(3): 351-361.

Gilbert, W. (1978). "Why genes in pieces?" Nature **271**(5645): 501.

Girard, C., C. L. Will, J. Peng, E. M. Makarov, B. Kastner, I. Lemm, H. Urlaub, K. Hartmuth and R. Luhrmann (2012). "Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion." Nat Commun **3**: 994.

Gonorazky, H., M. Liang, B. Cummings, M. Lek, J. Micallef, C. Hawkins, R. Basran, R. Cohn, M. D. Wilson, D. MacArthur, C. R. Marshall, P. N. Ray and J. J. Dowling (2016). "RNAseq analysis for the diagnosis of muscular dystrophy." Ann Clin Transl Neurol **3**(1): 55-60.

Gonzalez, I. L. (2005). "Barth syndrome: TAZ gene mutations, mRNAs, and evolution." Am J Med Genet A **134**(4): 409-414.

Graham, F. L., J. Smiley, W. C. Russell and R. Nairn (1977). "Characteristics of a human cell line transformed by DNA from human adenovirus type 5." J Gen Virol **36**(1): 59-74.

Grellscheid, S. N. and C. W. Smith (2006). "An apparent pseudo-exon acts both as an alternative exon that leads to nonsense-mediated decay and as a zero-length exon." Mol Cell Biol **26**(6): 2237-2246.

Griesenbeck, J., H. Tschochner and D. Grohmann (2017). "Structure and Function of RNA Polymerases and the Transcription Machineries." Subcell Biochem **83**: 225-270.

Gromak, N., S. West and N. J. Proudfoot (2006). "Pause sites promote transcriptional termination of mammalian RNA polymerase II." Mol Cell Biol **26**(10): 3986-3996.

Grosso, A. R., A. P. Leite, S. Carvalho, M. R. Matos, F. B. Martins, A. C. Vitor, J. M. Desterro, M. Carmo-Fonseca and S. F. de Almeida (2015). "Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma." Elife **4**.

Gruber, A. R., G. Martin, W. Keller and M. Zavolan (2014). "Means to an end: mechanisms of alternative polyadenylation of messenger RNA precursors." Wiley Interdiscip Rev RNA **5**(2): 183-196.

Gudipati, R. K., Z. Xu, A. Lebreton, B. Seraphin, L. M. Steinmetz, A. Jacquier and D. Libri (2012). "Extensive degradation of RNA precursors by the exosome in wild-type cells." Mol Cell **48**(3): 409-421.

Gurvich, O. L., T. M. Tuohy, M. T. Howard, R. S. Finkel, L. Medne, C. B. Anderson, R. B. Weiss, S. D. Wilton and K. M. Flanigan (2008). "DMD pseudoexon mutations: splicing efficiency, phenotype, and potential therapy." Ann Neurol **63**(1): 81-89.

Haimovich, G., C. M. Ecker, M. C. Dunagin, E. Eggan, A. Raj, J. E. Gerst and R. H. Singer (2017). "Intercellular mRNA trafficking via membrane nanotube-like extensions in mammalian cells." Proc Natl Acad Sci U S A **114**(46): E9873-E9882.

Hall, S. L. and R. A. Padgett (1996). "Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns." Science **271**(5256): 1716-1718.

Han, J., J. Xiong, D. Wang and X. D. Fu (2011). "Pre-mRNA splicing: where and when in the nucleus." Trends Cell Biol **21**(6): 336-343.

Han, Y. and Y. He (2016). "Eukaryotic transcription initiation machinery visualized at molecular level." Transcription **7**(5): 203-208.

Hang, J., R. Wan, C. Yan and Y. Shi (2015). "Structural basis of pre-mRNA splicing." Science **349**(6253): 1191-1198.

Hare, M. P. and S. R. Palumbi (2003). "High intron sequence conservation across three mammalian orders suggests functional constraints." Mol Biol Evol **20**(6): 969-978.

Harlen, K. M., K. L. Trotta, E. E. Smith, M. M. Mosaheb, S. M. Fuchs and L. S. Churchman (2016). "Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue." Cell Rep **15**(10): 2147-2158.

Hatton, A. R., V. Subramaniam and A. J. Lopez (1998). "Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions." Mol Cell **2**(6): 787-796.

He, H., S. Liyanarachchi, K. Akagi, R. Nagy, J. Li, R. C. Dietrich, W. Li, N. Sebastian, B. Wen, B. Xin, J. Singh, P. Yan, H. Alder, E. Haan, D. Wieczorek, B. Albrecht, E. Puffenberger, H. Wang, J. A. Westman, R. A. Padgett, D. E. Symer and A. de la Chapelle (2011). "Mutations in U4atac snRNA, a component of the minor spliceosome, in the developmental disorder MOPD I." Science **332**(6026): 238-240.

He, Y., C. Yan, J. Fang, C. Inouye, R. Tjian, I. Ivanov and E. Nogales (2016). "Near-atomic resolution visualization of human transcription promoter opening." Nature **533**(7603): 359-365.

He, Y., C. Yuan, L. Chen, M. Lei, L. Zellmer, H. Huang and D. J. Liao (2018). "Transcriptional-Readthrough RNAs Reflect the Phenomenon of "A Gene Contains Gene(s)" or "Gene(s) within a Gene" in the Human Genome, and Thus Are Not Chimeric RNAs." Genes (Basel) **9**(1).

Heidemann, M., C. Hintermair, K. Voss and D. Eick (2013). "Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription." Biochim Biophys Acta **1829**(1): 55-62.

Heinzen, E. L., D. Ge, K. D. Cronin, J. M. Maia, K. V. Shianna, W. N. Gabriel, K. A. Welsh-Bohmer, C. M. Hulette, T. N. Denny and D. B. Goldstein (2008). "Tissue-specific genetic control of splicing: implications for the study of complex traits." PLoS Biol **6**(12): e1.

Herzel, L. and K. M. Neugebauer (2015). "Quantification of co-transcriptional splicing from RNA-Seq data." Methods **85**: 36-43.

Higgs, D. R., S. E. Goodbourn, J. Lamb, J. B. Clegg, D. J. Weatherall and N. J. Proudfoot (1983). "Alpha-thalassaemia caused by a polyadenylation signal mutation." Nature **306**(5941): 398-400.

Hirschfeld, M., A. zur Hausen, H. Bettendorf, M. Jager and E. Stickeler (2009). "Alternative splicing of Cyr61 is regulated by hypoxia and significantly changed in breast cancer." Cancer Res **69**(5): 2082-2090.

Holbrook, J. A., G. Neu-Yilik, M. W. Hentze and A. E. Kulozik (2004). "Nonsense-mediated decay approaches the clinic." Nat Genet **36**(8): 801-808.

Hollander, D., S. Naftelberg, G. Lev-Maor, A. R. Kornblihtt and G. Ast (2016). "How Are Short Exons Flanked by Long Introns Defined and Committed to Splicing?" Trends Genet **32**(10): 596-606.

Houseley, J. and D. Tollervey (2009). "The many pathways of RNA degradation." Cell **136**(4): 763-776.

Hsiao, Y. H., J. H. Bahn, X. Lin, T. M. Chan, R. Wang and X. Xiao (2016). "Alternative splicing modulated by genetic variants demonstrates accelerated evolution regulated by highly conserved proteins." Genome Res **26**(4): 440-450.

Hube, F. and C. Francastel (2015). "Mammalian introns: when the junk generates molecular diversity." Int J Mol Sci **16**(3): 4429-4452.

Hughes, P., D. Marshall, Y. Reid, H. Parkes and C. Gelber (2007). "The costs of using unauthenticated, over-passaged cell lines: how much more data do we need?" Biotechniques **43**(5): 575, 577-578, 581-572 passim.

Huranova, M., I. Ivani, A. Benda, I. Poser, Y. Brody, M. Hof, Y. Shav-Tal, K. M. Neugebauer and D. Stanek (2010). "The differential interaction of snRNPs with pre-mRNA reveals splicing kinetics in living cells." J Cell Biol **191**(1): 75-86.

Hussain, T. and R. Mulherkar (2012). "Lymphoblastoid Cell lines: a Continuous in Vitro Source of Cells to Study Carcinogen Sensitivity and DNA Repair." Int J Mol Cell Med 1(2): 75-87.

Iasillo, C., M. Schmid, Y. Yahia, M. A. Maqbool, N. Descostes, E. Karadoulama, E. Bertrand, J. C. Andrau and T. H. Jensen (2017). "ARS2 is a general suppressor of pervasive transcription." Nucleic Acids Res 45(17): 10229-10241.

Ivings, L., K. V. Towns, M. A. Matin, C. Taylor, F. Ponchel, R. J. Grainger, R. S. Ramesar, D. A. Mackey and C. F. Inglehearn (2008). "Evaluation of splicing efficiency in lymphoblastoid cell lines from patients with splicing-factor retinitis pigmentosa." Mol Vis 14: 2357-2366.

Jaillon, O., K. Bouhouche, J. F. Gout, J. M. Aury, B. Noel, B. Saudemont, M. Nowacki, V. Serrano, B. M. Porcel, B. Segurens, A. Le Mouel, G. Lepere, V. Schachter, M. Betermier, J. Cohen, P. Wincker, L. Sperling, L. Duret and E. Meyer (2008). "Translational control of intron splicing in eukaryotes." Nature 451(7176): 359-362.

Jeck, W. R., J. A. Sorrentino, K. Wang, M. K. Slevin, C. E. Burd, J. Liu, W. F. Marzluff and N. E. Sharpless (2013). "Circular RNAs are abundant, conserved, and associated with ALU repeats." RNA 19(2): 141-157.

Jenal, M., R. Elkon, F. Loayza-Puch, G. van Haaften, U. Kuhn, F. M. Menzies, J. A. Oude Vrielink, A. J. Bos, J. Drost, K. Rooijers, D. C. Rubinsztein and R. Agami (2012). "The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites." Cell 149(3): 538-553.

Jha, H. C., Y. Pei and E. S. Robertson (2016). "Epstein-Barr Virus: Diseases Linked to Infection and Transformation." Front Microbiol 7: 1602.

Jiao, X., J. H. Chang, T. Kilic, L. Tong and M. Kiledjian (2013). "A mammalian pre-mRNA 5' end capping quality control mechanism and an unexpected link of capping to pre-mRNA processing." Mol Cell 50(1): 104-115.

Johnson, C., D. Primorac, M. McKinstry, J. McNeil, D. Rowe and J. B. Lawrence (2000). "Tracking COL1A1 RNA in osteogenesis imperfecta. splice-defective transcripts initiate transport from the gene but are retained within the SC35 domain." J Cell Biol 150(3): 417-432.

Johnston, J., R. I. Kelley, A. Feigenbaum, G. F. Cox, G. S. Iyer, V. L. Funanage and R. Proujansky (1997). "Mutation characterization and genotype-phenotype correlation in Barth syndrome." Am J Hum Genet 61(5): 1053-1058.

Jonkers, I., H. Kwak and J. T. Lis (2014). "Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons." Elife 3: e02407.

Juneau, K., M. Miranda, M. E. Hillenmeyer, C. Nislow and R. W. Davis (2006). "Introns regulate RNA and protein abundance in yeast." Genetics 174(1): 511-518.

Kaida, D. (2016). "The reciprocal regulation between splicing and 3'-end processing." Wiley Interdiscip Rev RNA 7(4): 499-511.

Kaida, D., M. G. Berg, I. Younis, M. Kasim, L. N. Singh, L. Wan and G. Dreyfuss (2010). "U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation." Nature 468(7324): 664-668.

Kelly, S., T. Georgomanolis, A. Zirkel, S. Diermeier, D. O'Reilly, S. Murphy, G. Langst, P. R. Cook and A. Papantonis (2015). "Splicing of many human genes involves sites embedded within introns." Nucleic Acids Res 43(9): 4721-4732.

Kenzelmann, M., S. Maertens, M. Hergenhahn, S. Kueffer, A. Hotz-Wagenblatt, L. Li, S. Wang, C. Ittrich, T. Lemberger, R. Arribas, S. Jonnakuty, M. C. Hollstein, W. Schmid, N. Gretz, H. J. Grone and G. Schutz (2007). "Microarray analysis of newly synthesized RNA in cells and animals." Proc Natl Acad Sci U S A 104(15): 6164-6169.

Keren-Shaul, H., G. Lev-Maor and G. Ast (2013). "Pre-mRNA splicing is a determinant of nucleosome organization." PLoS One 8(1): e53506.

Keren, H., G. Lev-Maor and G. Ast (2010). "Alternative splicing and evolution: diversification, exon definition and function." Nat Rev Genet 11(5): 345-355.

Kervestin, S. and A. Jacobson (2012). "NMD: a multifaceted response to premature translational termination." Nat Rev Mol Cell Biol 13(11): 700-712.

Khan, S. G., K. S. Oh, S. Emmert, K. Imoto, D. Tamura, J. J. Digiovanna, T. Shahlavi, N. Armstrong, C. C. Baker, M. Neuburg, C. Zalewski, C. Brewer, E. Wiggs, R. Schiffmann and K. H. Kraemer (2009). "XPC initiation codon mutation in xeroderma pigmentosum patients with and without neurological symptoms." DNA Repair (Amst) 8(1): 114-125.

Khan, S. G., K. S. Oh, T. Shahlavi, T. Ueda, D. B. Busch, H. Inui, S. Emmert, K. Imoto, V. Muniz-Medina, C. C. Baker, J. J. DiGiovanna, D. Schmidt, A. Khadavi, A. Metin, E. Gozukara, H. Slor, A. Sarasin and K. H. Kraemer (2006). "Reduced XPC DNA repair gene mRNA levels in clinically normal parents of xeroderma pigmentosum patients." Carcinogenesis 27(1): 84-94.

Khodor, Y. L., J. S. Menet, M. Tolan and M. Rosbash (2012). "Cotranscriptional splicing efficiency differs dramatically between Drosophila and mouse." RNA 18(12): 2174-2186.

Khodor, Y. L., J. Rodriguez, K. C. Abruzzi, C. H. Tang, M. T. Marr, 2nd and M. Rosbash (2011). "Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila." Genes Dev 25(23): 2502-2512.

Kilchert, C., S. Wittmann and L. Vasiljeva (2016). "The regulation and functions of the nuclear RNA exosome complex." Nat Rev Mol Cell Biol 17(4): 227-239.

Kirwin, S. M., A. Manolakos, S. S. Barnett and I. L. Gonzalez (2014). "Tafazzin splice variants and mutations in Barth syndrome." Mol Genet Metab 111(1): 26-32.

Krawczak, M., N. S. Thomas, B. Hundrieser, M. Mort, M. Wittig, J. Hampe and D. N. Cooper (2007). "Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing." Hum Mutat **28**(2): 150-158.

Krebs, A. R., D. Imanci, L. Hoerner, D. Gaidatzis, L. Burger and D. Schubeler (2017). "Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters." Mol Cell **67**(3): 411-422 e414.

Kumar-Sinha, C., S. Kalyana-Sundaram and A. M. Chinnaiyan (2012). "SLC45A3-ELK4 chimera in prostate cancer: spotlight on cis-splicing." Cancer Discov **2**(7): 582-585.

Kumar, S., J. E. Curran, D. C. Glahn and J. Blangero (2016). "Utility of Lymphoblastoid Cell Lines for Induced Pluripotent Stem Cell Generation." Stem Cells Int **2016**: 2349261.

Kurio, H., E. Murayama, T. Kaneko, Y. Shibata, T. Inai and H. Iida (2008). "Intron retention generates a novel isoform of CEACAM6 that may act as an adhesion molecule in the ectoplasmic specialization structures between spermatids and sertoli cells in rat testis." Biol Reprod **79**(6): 1062-1073.

Kurosaki, T. and L. E. Maquat (2016). "Nonsense-mediated mRNA decay in humans at a glance." J Cell Sci **129**(3): 461-467.

Kwak, H., N. J. Fuda, L. J. Core and J. T. Lis (2013). "Precise maps of RNA polymerase reveal how promoters direct initiation and pausing." Science **339**(6122): 950-953.

Kwak, H. and J. T. Lis (2013). "Control of transcriptional elongation." Annu Rev Genet **47**: 483-508.

Kwek, K. Y., S. Murphy, A. Furger, B. Thomas, W. O'Gorman, H. Kimura, N. J. Proudfoot and A. Akoulitchev (2002). "U1 snRNA associates with TFIIH and regulates transcriptional initiation." Nat Struct Biol **9**(11): 800-805.

Kyburz, A., A. Friedlein, H. Langen and W. Keller (2006). "Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3' end processing and splicing." Mol Cell **23**(2): 195-205.

Labno, A., R. Tomecki and A. Dziembowski (2016). "Cytoplasmic RNA decay pathways - Enzymes and mechanisms." Biochim Biophys Acta **1863**(12): 3125-3147.

Lane, A. N. and T. W. Fan (2015). "Regulation of mammalian nucleotide metabolism and biosynthesis." Nucleic Acids Res **43**(4): 2466-2485.

Lareau, L. F., M. Inada, R. E. Green, J. C. Wengrod and S. E. Brenner (2007). "Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements." Nature **446**(7138): 926-929.

Larson, D. R., R. H. Singer and D. Zenklusen (2009). "A single molecule view of gene expression." Trends Cell Biol **19**(11): 630-637.

Larson, D. R., D. Zenklusen, B. Wu, J. A. Chao and R. H. Singer (2011). "Real-time observation of transcription initiation and elongation on an endogenous yeast gene." Science 332(6028): 475-478.

Le Hir, H., J. Sauliere and Z. Wang (2016). "The exon junction complex as a node of post-transcriptional networks." Nat Rev Mol Cell Biol 17(1): 41-54.

Lee, E. S., A. Akef, K. Mahadevan and A. F. Palazzo (2015). "The consensus 5' splice site motif inhibits mRNA nuclear export." PLoS One 10(3): e0122743.

Liang, D. and J. E. Wilusz (2014). "Short intronic repeat sequences facilitate circular RNA production." Genes Dev 28(20): 2233-2247.

Liu, H. X., M. Zhang and A. R. Krainer (1998). "Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins." Genes Dev 12(13): 1998-2012.

Liu, X., D. A. Bushnell, D. A. Silva, X. Huang and R. D. Kornberg (2011). "Initiation complex structure and promoter proofreading." Science 333(6042): 633-637.

Liu, Z., I. Luyten, M. J. Bottomley, A. C. Messias, S. Houngninou-Molango, R. Sprangers, K. Zanier, A. Kramer and M. Sattler (2001). "Structural basis for recognition of the intron branch site RNA by splicing factor 1." Science 294(5544): 1098-1102.

Lopez-Bigas, N., B. Audit, C. Ouzounis, G. Parra and R. Guigo (2005). "Are splicing mutations the most frequent cause of hereditary disease?" FEBS Lett 579(9): 1900-1903.

Louhichi, A., A. Fourati and A. Rebai (2011). "IGD: a resource for intronless genes in the human genome." Gene 488(1-2): 35-40.

Lubas, M., P. R. Andersen, A. Schein, A. Dziembowski, G. Kudla and T. H. Jensen (2015). "The human nuclear exosome targeting complex is loaded onto newly synthesized RNA to direct early ribonucleolysis." Cell Rep 10(2): 178-192.

Lucas, B. A., E. Lavi, L. Shiue, H. Cho, S. Katzman, K. Miyoshi, M. C. Siomi, L. Carmel, M. Ares, Jr. and L. E. Maquat (2018). "Evidence for convergent evolution of SINE-directed Staufen-mediated mRNA decay." Proc Natl Acad Sci U S A 115(5): 968-973.

Luco, R. F., M. Allo, I. E. Schor, A. R. Kornblihtt and T. Misteli (2011). "Epigenetics in alternative pre-mRNA splicing." Cell 144(1): 16-26.

Mandel, C. R., S. Kaneko, H. Zhang, D. Gebauer, V. Vethantham, J. L. Manley and L. Tong (2006). "Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease." Nature 444(7121): 953-956.

Martin, R. M., J. Rino, C. Carvalho, T. Kirchhausen and M. Carmo-Fonseca (2013). "Live-cell visualization of pre-mRNA splicing with single-molecule sensitivity." Cell Rep 4(6): 1144-1155.

Martins, S. B., J. Rino, T. Carvalho, C. Carvalho, M. Yoshida, J. M. Klose, S. F. de Almeida and M. Carmo-Fonseca (2011). "Spliceosome assembly is coupled to RNA polymerase II dynamics at the 3' end of human genes." Nat Struct Mol Biol **18**(10): 1115-1123.

Mattick, J. S. (2001). "Non-coding RNAs: the architects of eukaryotic complexity." EMBO Rep **2**(11): 986-991.

Mattick, J. S. and M. J. Gagen (2001). "The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms." Mol Biol Evol **18**(9): 1611-1630.

Mauger, O., F. Lemoine and P. Scheiffele (2016). "Targeted Intron Retention and Excision for Rapid Gene Regulation in Response to Neuronal Activity." Neuron **92**(6): 1266-1278.

Maxwell, E. S. and M. J. Fournier (1995). "The small nucleolar RNAs." Annu Rev Biochem **64**: 897-934.

Mayer, A., J. di Iulio, S. Maleri, U. Eser, J. Vierstra, A. Reynolds, R. Sandstrom, J. A. Stamatoyannopoulos and L. S. Churchman (2015). "Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution." Cell **161**(3): 541-554.

Mayer, A., H. M. Landry and L. S. Churchman (2017). "Pause & go: from the discovery of RNA polymerase pausing to its functional implications." Curr Opin Cell Biol **46**: 72-80.

McCracken, S., N. Fong, K. Yankulov, S. Ballantyne, G. Pan, J. Greenblatt, S. D. Patterson, M. Wickens and D. L. Bentley (1997). "The C-terminal domain of RNA polymerase II couples mRNA processing to transcription." Nature **385**(6614): 357-361.

McGinty, R. K. and S. Tan (2015). "Nucleosome structure and function." Chem Rev **115**(6): 2255-2273.

McKenzie, R. W. and M. D. Brennan (1996). "The two small introns of the Drosophila affinidisjuncta Adh gene are required for normal transcription." Nucleic Acids Res **24**(18): 3635-3642.

Melvin, W. T., H. B. Milne, A. A. Slater, H. J. Allen and H. M. Keir (1978). "Incorporation of 6-thioguanosine and 4-thiouridine into RNA. Application to isolation of newly synthesised RNA by affinity chromatography." Eur J Biochem **92**(2): 373-379.

Mendell, J. T., N. A. Sharifi, J. L. Meyers, F. Martinez-Murillo and H. C. Dietz (2004). "Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise." Nat Genet **36**(10): 1073-1078.

Meola, N., M. Domanski, E. Karadoulama, Y. Chen, C. Gentil, D. Pultz, K. Vitting-Seerup, S. Lykke-Andersen, J. S. Andersen, A. Sandelin and T. H. Jensen (2016). "Identification of a Nuclear Exosome Decay Pathway for Processed Transcripts." Mol Cell **64**(3): 520-533.

Merendino, L., S. Guth, D. Bilbao, C. Martinez and J. Valcarcel (1999). "Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF35 and the 3' splice site AG." Nature 402(6763): 838-841.

Merico, D., M. Roifman, U. Braunschweig, R. K. Yuen, R. Alexandrova, A. Bates, B. Reid, T. Nalpathamkalam, Z. Wang, B. Thiruvahindrapuram, P. Gray, A. Kakakios, J. Peake, S. Hogarth, D. Manson, R. Buncic, S. L. Pereira, J. A. Herbrick, B. J. Blencowe, C. M. Roifman and S. W. Scherer (2015). "Compound heterozygous mutations in the noncoding RNU4ATAC cause Roifman Syndrome by disrupting minor intron splicing." Nat Commun 6: 8718.

Metze, S., V. A. Herzog, M. D. Ruepp and O. Muhlemann (2013). "Comparison of EJC-enhanced and EJC-independent NMD in human cells reveals two partially redundant degradation pathways." RNA 19(10): 1432-1448.

Meyer, L. R., A. S. Zweig, A. S. Hinrichs, D. Karolchik, R. M. Kuhn, M. Wong, C. A. Sloan, K. R. Rosenbloom, G. Roe, B. Rhead, B. J. Raney, A. Pohl, V. S. Malladi, C. H. Li, B. T. Lee, K. Learned, V. Kirkup, F. Hsu, S. Heitner, R. A. Harte, M. Haeussler, L. Guruvadoo, M. Goldman, B. M. Giardine, P. A. Fujita, T. R. Dreszer, M. Diekhans, M. S. Cline, H. Clawson, G. P. Barber, D. Haussler and W. J. Kent (2013). "The UCSC Genome Browser database: extensions and updates 2013." Nucleic Acids Res 41(Database issue): D64-69.

Miki, T. S. and H. Grosshans (2013). "The multifunctional RNase XRN2." Biochem Soc Trans 41(4): 825-830.

Miller, C., B. Schwalb, K. Maier, D. Schulz, S. Dumcke, B. Zacher, A. Mayer, J. Sydow, L. Marcinowski, L. Dolken, D. E. Martin, A. Tresch and P. Cramer (2011). "Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast." Mol Syst Biol 7: 458.

Miller, G. (1982). "Immortalization of human lymphocytes by Epstein-Barr virus." Yale J Biol Med 55(3-4): 305-310.

Miller, J. N. and D. A. Pearce (2014). "Nonsense-mediated decay in genetic disease: friend or foe?" Mutat Res Rev Mutat Res 762: 52-64.

Miller, M. R., K. J. Robinson, M. D. Cleary and C. Q. Doe (2009). "TU-tagging: cell type-specific RNA isolation from intact complex tissues." Nat Methods 6(6): 439-441.

Millevoi, S. and S. Vagner (2010). "Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation." Nucleic Acids Res 38(9): 2757-2774.

Misteli, T. and D. L. Spector (1999). "RNA polymerase II targets pre-mRNA splicing factors to transcription sites in vivo." Mol Cell 3(6): 697-705.

Morales, J. C., P. Richard, P. L. Patidar, E. A. Motea, T. T. Dang, J. L. Manley and D. A. Boothman (2016). "XRN2 Links Transcription Termination to DNA Damage and Replication Stress." PLoS Genet 12(7): e1006107.

Muhlemann, O. and T. H. Jensen (2012). "mRNP quality control goes regulatory." Trends Genet 28(2): 70-77.

Muhlemann, O. and J. Lykke-Andersen (2010). "How and where are nonsense mRNAs degraded in mammalian cells?" RNA Biol 7(1): 28-32.

Mullen, N. J. and D. H. Price (2017). "Hydrogen peroxide yields mechanistic insights into human mRNA capping enzyme function." PLoS One 12(10): e0186423.

Muniz, L., M. K. Deb, M. Aguirrebengoa, S. Lazorthes, D. Trouche and E. Nicolas (2017). "Control of Gene Expression in Senescence through Transcriptional Read-Through of Convergent Protein-Coding Genes." Cell Rep 21(9): 2433-2446.

Naftelberg, S., I. E. Schor, G. Ast and A. R. Kornblihtt (2015). "Regulation of alternative splicing through coupling with transcription and chromatin structure." Annu Rev Biochem 84: 165-198.

Nag, A. and J. A. Steitz (2012). "Tri-snRNP-associated proteins interact with subunits of the TRAMP and nuclear exosome complexes, linking RNA decay and pre-mRNA splicing." RNA Biol 9(3): 334-342.

Nagarajan, V. K., C. I. Jones, S. F. Newbury and P. J. Green (2013). "XRN 5'-->3' exoribonucleases: structure, mechanisms and functions." Biochim Biophys Acta 1829(6-7): 590-603.

Nagy, E. and L. E. Maquat (1998). "A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance." Trends Biochem Sci 23(6): 198-199.

Naro, C., A. Jolly, S. Di Persio, P. Bielli, N. Setterblad, A. J. Alberdi, E. Vicini, R. Geremia, P. De la Grange and C. Sette (2017). "An Orchestrated Intron Retention Program in Meiosis Controls Timely Usage of Transcripts during Germ Cell Differentiation." Dev Cell 41(1): 82-93 e84.

Nayak, G., S. I. Lee, R. Yousaf, S. E. Edelmann, C. Trincot, C. M. Van Itallie, G. P. Sinha, M. Rafeeq, S. M. Jones, I. A. Belyantseva, J. M. Anderson, A. Forge, G. I. Frolenkov and S. Riazuddin (2013). "Tricellulin deficiency affects tight junction architecture and cochlear hair cells." J Clin Invest 123(9): 4036-4049.

Nechaev, S. and K. Adelman (2011). "Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation." Biochim Biophys Acta 1809(1): 34-45.

Nickless, A., J. M. Bailis and Z. You (2017). "Control of gene expression through the nonsense-mediated RNA decay pathway." Cell Biosci 7: 26.

Noah, D. L., K. Y. Twu and R. M. Krug (2003). "Cellular antiviral responses against influenza A virus are countered at the posttranscriptional level by the viral NS1A protein

via its binding to a cellular protein required for the 3' end processing of cellular pre-mRNAS." Virology 307(2): 386-395.

Nogales, E., R. K. Louder and Y. He (2017). "Structural Insights into the Eukaryotic Transcription Initiation Machinery." Annu Rev Biophys 46: 59-83.

Nogues, G., S. Kadener, P. Cramer, M. de la Mata, J. P. Fededa, M. Blaustein, A. Srebrow and A. R. Kornblihtt (2003). "Control of alternative pre-mRNA splicing by RNA Pol II elongation: faster is not always better." IUBMB Life 55(4-5): 235-241.

Nojima, T., M. Dienstbier, S. Murphy, N. J. Proudfoot and M. J. Dye (2013). "Definition of RNA polymerase II CoTC terminator elements in the human genome." Cell Rep 3(4): 1080-1092.

Nojima, T., T. Gomes, A. R. Grosso, H. Kimura, M. J. Dye, S. Dhir, M. Carmo-Fonseca and N. J. Proudfoot (2015). "Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing." Cell 161(3): 526-540.

Nojima, T., T. Gomes, A. R. F. Grosso, H. Kimura, M. J. Dye, S. Dhir, M. Carmo-Fonseca and N. J. Proudfoot (2015). "Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing." Cell 161(3): 526-540.

Okamura, K., J. W. Hagen, H. Duan, D. M. Tyler and E. C. Lai (2007). "The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila." Cell 130(1): 89-100.

Orkin, S. H., T. C. Cheng, S. E. Antonarakis and H. H. Kazazian, Jr. (1985). "Thalassemia due to a mutation in the cleavage-polyadenylation signal of the human beta-globin gene." EMBO J 4(2): 453-456.

Ottens, F., V. Boehm, C. R. Sibley, J. Ule and N. H. Gehring (2017). "Transcript-specific characteristics determine the contribution of endo- and exonucleolytic decay pathways during the degradation of nonsense-mediated decay substrates." RNA 23(8): 1224-1236.

Ottens, F. and N. H. Gehring (2016). "Physiological and pathophysiological role of nonsense-mediated mRNA decay." Pflugers Arch 468(6): 1013-1028.

Padgett, R. A. (2012). "New connections between splicing and human disease." Trends Genet 28(4): 147-154.

Palangat, M. and D. R. Larson (2016). "Single-gene dual-color reporter cell line to analyze RNA synthesis in vivo." Methods 103: 77-85.

Palazzo, A. F. and E. S. Lee (2015). "Non-coding RNA: what is functional and what is junk?" Front Genet 6: 2.

Pandya-Jones, A. and D. L. Black (2009). "Co-transcriptional splicing of constitutive and alternative exons." RNA 15(10): 1896-1908.

Papasaikas, P. and J. Valcarcel (2016). "The Spliceosome: The Ultimate RNA Chaperone and Sculptor." Trends Biochem Sci **41**(1): 33-45.

Park, E. and L. E. Maquat (2013). "Staufen-mediated mRNA decay." Wiley Interdiscip Rev RNA **4**(4): 423-435.

Park, S. G., S. Hannenhalli and S. S. Choi (2014). "Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals." BMC Genomics **15**: 526.

Patel, A. A. and J. A. Steitz (2003). "Splicing double: insights from the second spliceosome." Nat Rev Mol Cell Biol **4**(12): 960-970.

Pedrotti, S. and T. A. Cooper (2014). "In Brief: (mis)splicing in disease." J Pathol **233**(1): 1-3.

Pickrell, J. K., A. A. Pai, Y. Gilad and J. K. Pritchard (2010). "Noisy splicing drives mRNA isoform diversity in human cells." PLoS Genet **6**(12): e1001236.

Popp, M. W. and L. E. Maquat (2013). "Organizing principles of mammalian nonsense-mediated mRNA decay." Annu Rev Genet **47**: 139-165.

Porrua, O. and D. Libri (2013). "RNA quality control in the nucleus: the Angels' share of RNA." Biochim Biophys Acta **1829**(6-7): 604-611.

Poss, Z. C., C. C. Ebmeier and D. J. Taatjes (2013). "The Mediator complex and transcription regulation." Crit Rev Biochem Mol Biol **48**(6): 575-608.

Proudfoot, N. J. (2011). "Ending the message: poly(A) signals then and now." Genes Dev **25**(17): 1770-1782.

Proudfoot, N. J. (2016). "Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut." Science **352**(6291): aad9926.

Proudfoot, N. J. and G. G. Brownlee (1976). "3' non-coding region sequences in eukaryotic messenger RNA." Nature **263**(5574): 211-214.

Proudfoot, N. J., A. Furger and M. J. Dye (2002). "Integrating mRNA processing with transcription." Cell **108**(4): 501-512.

Rabani, M., R. Raychowdhury, M. Jovanovic, M. Rooney, D. J. Stumpo, A. Pauli, N. Hacohen, A. F. Schier, P. J. Blackshear, N. Friedman, I. Amit and A. Regev (2014). "High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies." Cell **159**(7): 1698-1710.

Rajavel, K. S. and E. F. Neufeld (2001). "Nonsense-mediated decay of human HEXA mRNA." Mol Cell Biol **21**(16): 5512-5519.

Riazuddin, S., Z. M. Ahmed, A. S. Fanning, A. Lagziel, S. Kitajiri, K. Ramzan, S. N. Khan, P. Chattaraj, P. L. Friedman, J. M. Anderson, I. A. Belyantseva, A. Forge, S. Riazuddin and T.

B. Friedman (2006). "Tricellulin is a tight-junction protein necessary for hearing." Am J Hum Genet 79(6): 1040-1051.

Richmond, T. J. and C. A. Davey (2003). "The structure of DNA in the nucleosome core." Nature 423(6936): 145-150.

Rigo, F. and H. G. Martinson (2008). "Functional coupling of last-intron splicing and 3'-end processing to transcription in vitro: the poly(A) signal couples to splicing before committing to cleavage." Mol Cell Biol 28(2): 849-862.

Rino, J., A. C. de Jesus and M. Carmo-Fonseca (2016). "STaQTool: Spot tracking and quantification tool for monitoring splicing of single pre-mRNA molecules in living cells." Methods 98: 143-149.

Rino, J., A. C. de Jesus and M. Carmo-Fonseca (2017). "Quantitative Image Analysis of Single-Molecule mRNA Dynamics in Living Cells." Methods Mol Biol 1563: 229-242.

Roca, X., M. Akerman, H. Gaus, A. Berdeja, C. F. Bennett and A. R. Krainer (2012). "Widespread recognition of 5' splice sites by noncanonical base-pairing to U1 snRNA involving bulged nucleotides." Genes Dev 26(10): 1098-1109.

Roca, X., A. R. Krainer and I. C. Eperon (2013). "Pick one, but be quick: 5' splice sites and the problems of too many choices." Genes Dev 27(2): 129-144.

Roca, X., R. Sachidanandam and A. R. Krainer (2003). "Intrinsic differences between authentic and cryptic 5' splice sites." Nucleic Acids Res 31(21): 6321-6333.

Rogozin, I. B., Y. I. Wolf, A. V. Sorokin, B. G. Mirkin and E. V. Koonin (2003). "Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution." Curr Biol 13(17): 1512-1517.

Romano, M., E. Buratti and D. Baralle (2013). "Role of pseudoexons and pseudointrons in human cancer." Int J Cell Biol 2013: 810572.

Roy, S. W. and M. Irimia (2008). "Intron mis-splicing: no alternative?" Genome Biol 9(2): 208.

Ruskin, B. and M. R. Green (1985). "An RNA processing activity that debranches RNA lariats." Science 229(4709): 135-140.

Rutkowski, A. J., F. Erhard, A. L'Hernault, T. Bonfert, M. Schilhabel, C. Crump, P. Rosenstiel, S. Efstathiou, R. Zimmer, C. C. Friedel and L. Dolken (2015). "Widespread disruption of host transcription termination in HSV-1 infection." Nat Commun 6: 7126.

Sagai, T., M. Hosoya, Y. Mizushina, M. Tamura and T. Shiroishi (2005). "Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb." Development 132(4): 797-803.

Sainsbury, S., J. Niesser and P. Cramer (2013). "Structure and function of the initially transcribing RNA polymerase II-TFIIB complex." Nature 493(7432): 437-440.

Salzman, J. (2016). "Circular RNA Expression: Its Potential Regulation and Function." Trends Genet **32**(5): 309-316.

Sandberg, R., J. R. Neilson, A. Sarma, P. A. Sharp and C. B. Burge (2008). "Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites." Science **320**(5883): 1643-1647.

Schmid, M. and T. H. Jensen (2010). "Nuclear quality control of RNA polymerase II transcripts." Wiley Interdiscip Rev RNA **1**(3): 474-485.

Schmidt, U., E. Basyuk, M. C. Robert, M. Yoshida, J. P. Villemin, D. Auboeuf, S. Aitken and E. Bertrand (2011). "Real-time imaging of cotranscriptional splicing reveals a kinetic model that reduces noise: implications for alternative splicing regulation." J Cell Biol **193**(5): 819-829.

Schmittgen, T. D. and K. J. Livak (2008). "Analyzing real-time PCR data by the comparative C(T) method." Nat Protoc **3**(6): 1101-1108.

Schneider-Poetsch, T., J. Ju, D. E. Eyler, Y. Dang, S. Bhat, W. C. Merrick, R. Green, B. Shen and J. O. Liu (2010). "Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin." Nat Chem Biol **6**(3): 209-217.

Schwalb, B., M. Michel, B. Zacher, K. Fruhauf, C. Demel, A. Tresch, J. Gagneur and P. Cramer (2016). "TT-seq maps the human transient transcriptome." Science **352**(6290): 1225-1228.

Scott, R. S. (2017). "Epstein-Barr virus: a master epigenetic manipulator." Curr Opin Virol **26**: 74-80.

Scotti, M. M. and M. S. Swanson (2016). "RNA mis-splicing in disease." Nat Rev Genet **17**(1): 19-32.

Seraphin, B. and M. Rosbash (1989). "Identification of functional U1 snRNA-pre-mRNA complexes committed to spliceosome assembly and splicing." Cell **59**(2): 349-358.

Shabalina, S. A., A. Y. Ogurtsov, A. N. Spiridonov, P. S. Novichkov, N. A. Spiridonov and E. V. Koonin (2010). "Distinct patterns of expression and evolution of intronless and intron-containing mammalian genes." Mol Biol Evol **27**(8): 1745-1749.

Sharp, P. A., M. M. Konarksa, P. J. Grabowski, A. I. Lamond, R. Marciniak and S. R. Seiler (1987). "Splicing of messenger RNA precursors." Cold Spring Harb Symp Quant Biol **52**: 277-285.

Shav-Tal, Y., X. Darzacq, S. M. Shenoy, D. Fusco, S. M. Janicki, D. L. Spector and R. H. Singer (2004). "Dynamics of single mRNPs in nuclei of living cells." Science **304**(5678): 1797-1800.

Shi, Y., D. C. Di Giammartino, D. Taylor, A. Sarkeshik, W. J. Rice, J. R. Yates, 3rd, J. Frank and J. L. Manley (2009). "Molecular architecture of the human pre-mRNA 3' processing complex." Mol Cell **33**(3): 365-376.

Shi, Y. and J. L. Manley (2015). "The end of the message: multiple protein-RNA interactions define the mRNA polyadenylation site." Genes Dev 29(9): 889-897.

Shukla, S., E. Kavak, M. Gregory, M. Imashimizu, B. Shutinoski, M. Kashlev, P. Oberdoerffer, R. Sandberg and S. Oberdoerffer (2011). "CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing." Nature 479(7371): 74-79.

Sibley, C. R., L. Blazquez and J. Ule (2016). "Lessons from non-canonical splicing." Nat Rev Genet 17(7): 407-421.

Sibley, C. R., W. Emmett, L. Blazquez, A. Faro, N. Haberman, M. Briese, D. Trabzuni, M. Ryten, M. E. Weale, J. Hardy, M. Modic, T. Curk, S. W. Wilson, V. Plagnol and J. Ule (2015). "Recursive splicing in long vertebrate genes." Nature 521(7552): 371-375.

Sie, L., S. Loong and E. K. Tan (2009). "Utility of lymphoblastoid cell lines." J Neurosci Res 87(9): 1953-1959.

Singh, J. and R. A. Padgett (2009). "Rates of in situ transcription and splicing in large human genes." Nat Struct Mol Biol 16(11): 1128-1133.

Singh, R. K. and T. A. Cooper (2012). "Pre-mRNA splicing in disease and therapeutics." Trends Mol Med 18(8): 472-482.

Sironi, M., G. Menozzi, L. Riva, R. Cagliani, G. P. Comi, N. Bresolin, R. Giorda and U. Pozzoli (2004). "Silencer elements as possible inhibitors of pseudoexon splicing." Nucleic Acids Res 32(5): 1783-1791.

Sisodia, S. S., B. Sollner-Webb and D. W. Cleveland (1987). "Specificity of RNA maturation pathways: RNAs transcribed by RNA polymerase III are not substrates for splicing or polyadenylation." Mol Cell Biol 7(10): 3602-3612.

Smale, S. T. and R. Tjian (1985). "Transcription of herpes simplex virus tk sequences under the control of wild-type and mutant human RNA polymerase I promoters." Mol Cell Biol 5(2): 352-362.

Soldner, F. and R. Jaenisch (2012). "Medicine. iPSC disease modeling." Science 338(6111): 1155-1156.

Spritz, R. A., P. Jagadeeswaran, P. V. Choudary, P. A. Biro, J. T. Elder, J. K. deRiel, J. L. Manley, M. L. Gefter, B. G. Forget and S. M. Weissman (1981). "Base substitution in an intervening sequence of a beta+-thalassemic human globin gene." Proc Natl Acad Sci U S A 78(4): 2455-2459.

Stadhouders, R., A. van den Heuvel, P. Kolovos, R. Jorna, K. Leslie, F. Grosveld and E. Soler (2012). "Transcription regulation by distal enhancers: who's in the loop?" Transcription 3(4): 181-186.

Stenson, P. D., M. Mort, E. V. Ball, K. Shaw, A. Phillips and D. N. Cooper (2014). "The Human Gene Mutation Database: building a comprehensive mutation repository for

clinical and molecular genetics, diagnostic testing and personalized genomic medicine." Hum Genet **133**(1): 1-9.

Stepanenko, A. A. and V. V. Dmitrenko (2015). "HEK293 in cell biology and cancer research: phenotype, karyotype, tumorigenicity, and stress-induced genome-phenotype evolution." Gene **569**(2): 182-190.

Sterne-Weiler, T. and J. R. Sanford (2014). "Exon identity crisis: disease-causing mutations that disrupt the splicing code." Genome Biol **15**(1): 201.

Stoecklin, G. and O. Muhlemann (2013). "RNA decay mechanisms: specificity through diversity." Biochim Biophys Acta **1829**(6-7): 487-490.

Sugimoto, M., H. Tahara, T. Ide and Y. Furuichi (2004). "Steps involved in immortalization and tumorigenesis in human B-lymphoblastoid cell lines transformed by Epstein-Barr virus." Cancer Res **64**(10): 3361-3364.

Sun, Y., Y. Zhang, K. Hamilton, J. L. Manley, Y. Shi, T. Walz and L. Tong (2017). "Molecular basis for the recognition of the human AAUAAA polyadenylation signal." Proc Natl Acad Sci U S A.

Tani, H. and N. Akimitsu (2012). "Genome-wide technology for determining RNA stability in mammalian cells: historical perspective and recent advantages based on modified nucleotide labeling." RNA Biol **9**(10): 1233-1238.

Tantale, K., F. Mueller, A. Kozulic-Pirher, A. Lesne, J. M. Victor, M. C. Robert, S. Capozi, R. Chouaib, V. Backer, J. Mateos-Langerak, X. Darzacq, C. Zimmer, E. Basyuk and E. Bertrand (2016). "A single-molecule view of transcription reveals convoys of RNA polymerases and multi-scale bursting." Nat Commun **7**: 12248.

Tarn, W. Y. and J. A. Steitz (1996). "Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns." Science **273**(5283): 1824-1832.

Tarn, W. Y. and J. A. Steitz (1996). "A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro." Cell **84**(5): 801-811.

Tian, B. and J. L. Manley (2017). "Alternative polyadenylation of mRNA precursors." Nat Rev Mol Cell Biol **18**(1): 18-30.

Tilgner, H., D. G. Knowles, R. Johnson, C. A. Davis, S. Chakrabortty, S. Djebali, J. Curado, M. Snyder, T. R. Gingeras and R. Guigo (2012). "Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs." Genome Res **22**(9): 1616-1625.

Tilgner, H., C. Nikolaou, S. Althammer, M. Sammeth, M. Beato, J. Valcarcel and R. Guigo (2009). "Nucleosome positioning as a determinant of exon recognition." Nat Struct Mol Biol **16**(9): 996-1001.

Toma, K. G., I. Rebbapragada, S. Durand and J. Lykke-Andersen (2015). "Identification of elements in human long 3' UTRs that inhibit nonsense-mediated decay." RNA **21**(5): 887-897.

Tosato, G. and J. I. Cohen (2007). "Generation of Epstein-Barr Virus (EBV)-immortalized B cell lines." Curr Protoc Immunol **Chapter 7**: Unit 7 22.

Trabelsi, M., C. Beugnet, N. Deburgrave, V. Commere, L. Orhant, F. Leturcq and J. Chelly (2014). "When a mid-intronic variation of DMD gene creates an ESE site." Neuromuscul Disord **24**(12): 1111-1117.

Treisman, R., S. H. Orkin and T. Maniatis (1983). "Specific transcription and RNA splicing defects in five cloned beta-thalassaemia genes." Nature **302**(5909): 591-596.

Tsunemoto, R. K., K. T. Eade, J. W. Blanchard and K. K. Baldwin (2015). "Forward engineering neuronal diversity using direct reprogramming." EMBO J **34**(11): 1445-1455.

Tycowski, K. T., M. D. Shu and J. A. Steitz (1996). "A mammalian gene with introns instead of exons generating stable RNA products." Nature **379**(6564): 464-466.

Urbanek, M. O., P. Galka-Marciniak, M. Olejniczak and W. J. Krzyzosiak (2014). "RNA imaging in living cells - methods and applications." RNA Biol **11**(8): 1083-1095.

Ustianenko, D., J. Pasulka, Z. Feketova, L. Bednarik, D. Zigackova, A. Fortova, M. Zavolan and S. Vanacova (2016). "TUT-DIS3L2 is a mammalian surveillance pathway for aberrant structured non-coding RNAs." EMBO J **35**(20): 2179-2191.

Valen, E., P. Preker, P. R. Andersen, X. Zhao, Y. Chen, C. Ender, A. Dueck, G. Meister, A. Sandelin and T. H. Jensen (2011). "Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes." Nat Struct Mol Biol **18**(9): 1075-1082.

Vandenbroucke, II, J. Vandesompele, A. D. Paepe and L. Messiaen (2001). "Quantification of splice variants using real-time PCR." Nucleic Acids Res **29**(13): E68-68.

Varley, K. E., J. Gertz, B. S. Roberts, N. S. Davis, K. M. Bowling, M. K. Kirby, A. S. Nesmith, P. G. Oliver, W. E. Grizzle, A. Forero, D. J. Buchsbaum, A. F. LoBuglio and R. M. Myers (2014). "Recurrent read-through fusion transcripts in breast cancer." Breast Cancer Res Treat **146**(2): 287-297.

Vaz-Drago, R., N. Custodio and M. Carmo-Fonseca (2017). "Deep intronic mutations and human disease." Hum Genet **136**(9): 1093-1111.

Venkatesh, S. and J. L. Workman (2015). "Histone exchange, chromatin structure and the regulation of transcription." Nat Rev Mol Cell Biol **16**(3): 178-189.

Verbeeren, J., B. Verma, E. H. Niemela, K. Yap, E. V. Makeyev and M. J. Frilander (2017). "Alternative exon definition events control the choice between nuclear retention and cytoplasmic export of U11/U12-65K mRNA." PLoS Genet **13**(5): e1006824.

Vilborg, A., N. Sabath, Y. Wiesel, J. Nathans, F. Levy-Adam, T. A. Yario, J. A. Steitz and R. Shalgi (2017). "Comparative analysis reveals genomic features of stress-induced transcriptional readthrough." Proc Natl Acad Sci U S A **114**(40): E8362-E8371.

Vorechovsky, I. (2006). "Aberrant 3' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization." Nucleic Acids Res **34**(16): 4630-4641.

Wahl, M. C., C. L. Will and R. Luhrmann (2009). "The spliceosome: design principles of a dynamic RNP machine." Cell **136**(4): 701-718.

Wang, Z., M. E. Rolish, G. Yeo, V. Tung, M. Mawson and C. B. Burge (2004). "Systematic identification and analysis of exonic splicing silencers." Cell **119**(6): 831-845.

Wegener, M. and M. Muller-McNicoll (2017). "Nuclear retention of mRNAs - quality control, gene regulation and human disease." Semin Cell Dev Biol.

West, S., N. Gromak and N. J. Proudfoot (2004). "Human 5' --> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites." Nature **432**(7016): 522-525.

West, S., N. J. Proudfoot and M. J. Dye (2008). "Molecular dissection of mammalian RNA polymerase II transcriptional termination." Mol Cell **29**(5): 600-610.

Will, C. L. and R. Luhrmann (2011). "Spliceosome structure and function." Cold Spring Harb Perspect Biol **3**(7).

Will, C. L., C. Schneider, R. Reed and R. Luhrmann (1999). "Identification of both shared and distinct proteins in the major and minor spliceosomes." Science **284**(5422): 2003-2005.

Williams, L. R., L. L. Quinn, M. Rowe and J. Zuo (2015). "Induction of the Lytic Cycle Sensitizes Epstein-Barr Virus-Infected B Cells to NK Cell Killing That Is Counteracted by Virus-Mediated NK Cell Evasion Mechanisms in the Late Lytic Cycle." J Virol **90**(2): 947-958.

Wilusz, J. E. (2015). "Repetitive elements regulate circular RNA biogenesis." Mob Genet Elements **5**(3): 1-7.

Winczura, K., M. Schmid, C. Iasillo, K. R. Molloy, L. M. Harder, J. S. Andersen, J. LaCava and T. H. Jensen (2018). "Characterizing ZC3H18, a Multi-domain Protein at the Interface of RNA Production and Destruction Decisions." Cell Rep **22**(1): 44-58.

Windhager, L., T. Bonfert, K. Burger, Z. Ruzsics, S. Krebs, S. Kaufmann, G. Malterer, A. L'Hernault, M. Schilhabel, S. Schreiber, P. Rosenstiel, R. Zimmer, D. Eick, C. C. Friedel and L. Dolken (2012). "Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution." Genome Res **22**(10): 2031-2042.

Witten, J. T. and J. Ule (2011). "Understanding splicing regulation through RNA splicing maps." Trends Genet **27**(3): 89-97.

Wong, J. J., W. Ritchie, O. A. Ebner, M. Selbach, J. W. Wong, Y. Huang, D. Gao, N. Pinello, M. Gonzalez, K. Baidya, A. Thoeng, T. L. Khoo, C. G. Bailey, J. Holst and J. E. Rasko (2013). "Orchestrated intron retention regulates normal granulocyte differentiation." Cell **154**(3): 583-595.

Workman, E., A. Veith and D. J. Battle (2014). "U1A regulates 3' processing of the survival motor neuron mRNA." J Biol Chem **289**(6): 3703-3712.

Wu, S., C. M. Romfo, T. W. Nilsen and M. R. Green (1999). "Functional recognition of the 3' splice site AG by the splicing factor U2AF35." Nature **402**(6763): 832-835.

Wuarin, J. and U. Schibler (1994). "Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing." Mol Cell Biol **14**(11): 7219-7225.

Xiong, H. Y., B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, Q. Morris, Y. Barash, A. R. Krainer, N. Jojic, S. W. Scherer, B. J. Blencowe and B. J. Frey (2015). "RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease." Science **347**(6218): 1254806.

Xu, Q., D. Walker, A. Bernardo, J. Brodbeck, M. E. Balestra and Y. Huang (2008). "Intron-3 retention/splicing controls neuronal expression of apolipoprotein E in the CNS." J Neurosci **28**(6): 1452-1459.

Yap, K., Z. Q. Lim, P. Khandelia, B. Friedman and E. V. Makeyev (2012). "Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention." Genes Dev **26**(11): 1209-1223.

Yap, K. and E. V. Makeyev (2013). "Regulation of gene expression in mammalian nervous system through alternative pre-mRNA splicing coupled with RNA quality control mechanisms." Mol Cell Neurosci **56**: 420-428.

Yasuda, H., C. D. Oh, D. Chen, B. de Crombrugghe and J. H. Kim (2017). "A Novel Regulatory Mechanism of Type II Collagen Expression via a SOX9-dependent Enhancer in Intron 6." J Biol Chem **292**(2): 528-538.

Yoon, Y. J., B. Wu, A. R. Buxbaum, S. Das, A. Tsai, B. P. English, J. B. Grimm, L. D. Lavis and R. H. Singer (2016). "Glutamate-induced RNA localization and translation in neurons." Proc Natl Acad Sci U S A **113**(44): E6877-E6886.

Yu, Y., P. A. Maroney, J. A. Denker, X. H. Zhang, O. Dybkov, R. Luhrmann, E. Jankowsky, L. A. Chasin and T. W. Nilsen (2008). "Dynamic regulation of alternative splicing by silencers that modulate 5' splice site competition." Cell **135**(7): 1224-1236.

Yunger, S., L. Rosenfeld, Y. Garini and Y. Shav-Tal (2010). "Single-allele analysis of transcription kinetics in living mammalian cells." Nat Methods 7(8): 631-633.

Zamore, P. D., J. G. Patton and M. R. Green (1992). "Cloning and domain structure of the mammalian splicing factor U2AF." Nature 355(6361): 609-614.

Zhang, H., F. Rigo and H. G. Martinson (2015). "Poly(A) Signal-Dependent Transcription Termination Occurs through a Conformational Change Mechanism that Does Not Require Cleavage at the Poly(A) Site." Mol Cell 59(3): 437-448.

Zhang, X. H. and L. A. Chasin (2004). "Computational definition of sequence motifs governing constitutive exon splicing." Genes Dev 18(11): 1241-1250.

Zhang, Y., X. O. Zhang, T. Chen, J. F. Xiang, Q. F. Yin, Y. H. Xing, S. Zhu, L. Yang and L. L. Chen (2013). "Circular intronic long noncoding RNAs." Mol Cell 51(6): 792-806.

Zheng, C. L., Y. S. Kwon, H. R. Li, K. Zhang, G. Coutinho-Mansfield, C. Yang, T. M. Nair, M. Gribskov and X. D. Fu (2005). "MAASE: an alternative splicing database designed for supporting splicing microarray applications." RNA 11(12): 1767-1776.

Zhou, H., S. C. Schmidt, S. Jiang, B. Willox, K. Bernhardt, J. Liang, E. C. Johannsen, P. Kharchenko, B. E. Gewurz, E. Kieff and B. Zhao (2015). "Epstein-Barr virus oncoprotein super-enhancers control B cell growth." Cell Host Microbe 17(2): 205-216.

Zinder, J. C. and C. D. Lima (2017). "Targeting RNA for processing or destruction by the eukaryotic RNA exosome and its cofactors." Genes Dev 31(2): 88-100.

Zorio, D. A. and T. Blumenthal (1999). "Both subunits of U2AF recognize the 3' splice site in Caenorhabditis elegans." Nature 402(6763): 835-838.

Annexes

Human Molecular Genetics, 2015, Vol. 24, No. 10 2784–2795

doi: 10.1093/hmg/ddv039
Advance Access Publication Date: 4 February 2015
Original Article

ORIGINAL ARTICLE

Transcription-coupled RNA surveillance in human genetic diseases caused by splice site mutations

Rita Vaz-Drago, Marco T. Pinheiro†, Sandra Martins, Francisco J. Enguita, Maria Carmo-Fonseca* and Noélia Custódio*

Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisboa 1649-028, Portugal

*To whom correspondence should be addressed at: Instituto de Medicina Molecular, Faculdade de Medicina, Av. Prof. Egas Moniz, 1649-028 Lisboa, Portugal. Tel: +351 21 7999411; Fax: +351 21 7999412; Email: carmo.fonseca@medicina.ulisboa.pt (M.C-F.); noelia.custodio@medicina.ulisboa.pt (N.C.)

Abstract

Current estimates indicate that approximately one-third of all disease-causing mutations are expected to disrupt splicing. Abnormal splicing often leads to disruption of the reading frame with introduction of a premature termination codon (PTC) that targets the mRNA for degradation in the cytoplasm by nonsense mediated decay (NMD). In addition to NMD there are RNA surveillance mechanisms that act in the nucleus while transcripts are still associated with the chromatin template. However, the significance of nuclear RNA quality control in the context of human genetic diseases is unknown. Here we used patient-derived lymphoblastoid cell lines as disease models to address how biogenesis of mRNAs is affected by splice site mutations. We observed that most of the mutations analyzed introduce PTCs and trigger mRNA degradation in the cytoplasm. However, for some mutant transcripts, RNA levels associated with chromatin were found down-regulated. Quantification of nascent transcripts further revealed that a subset of genes containing splicing mutations (SM) have reduced transcriptional activity. Following treatment with the translation inhibitor cycloheximide the cytoplasmic levels of mutant RNAs increased, while the levels of chromatin-associated transcripts remained unaltered. These results suggest that transcription-coupled surveillance mechanisms operate independently from NMD to reduce cellular levels of abnormal RNAs caused by SM.

Introduction

Recent studies indicate that an unexpected high fraction of human diseases are caused by mutations that disrupt splicing. Approximately 15% of disease-causing mutations in the Human Gene Mutation Database (1) alter the consensus splice site sequences and another 22% are predicted to affect splicing enhancer or inhibitory motifs (2). Thus, altogether approximately one-third of all disease-causing mutations are assumed to affect splicing. The phenotypic outcome of these mutations can be summarized in three distinct categories: (1) Constitutive exon skipping or intron retention; (2) altered inclusion/ exclusion ratio of alternative exons; and (3) activation of cryptic splice sites, resulting in inclusion/ exclusion of sequences in a spliced mRNA. Most often the final result is gene loss of function due to either synthesis of a non-functional protein or disruption of the reading frame that introduces a premature termination codon (PTC) and targets the mRNA for degradation by nonsense mediated decay (NMD) (3).

Although NMD is a highly efficient post-transcriptional quality control mechanism that detects and destroys aberrant mRNAs containing PTCs (4), additional surveillance pathways ensure transcriptome fidelity. In particular, there is growing evidence indicating that cells can co-transcriptionally initiate the removal of abnormally processed pre-mRNAs as well as down-regulate transcription when pre-mRNA processing repeatedly fails (5). To date, several studies have characterized co-transcriptional RNA surveillance pathways in yeast and mammalian cells. In yeast, mutations in the splicing machinery did not result in steady-state accumulation of unspliced pre-mRNA unless the

† Present address: Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, UK.

Received: January 9, 2015. **Revised:** January 9, 2015. **Accepted:** January 31, 2015

exosome was rendered defective (6). These early experiments suggested that aberrantly processed transcripts are recognized and targeted for degradation by the exosome in the nucleus. A potential player in this process is the poly(A) binding protein Nab2p/Pab2p, which binds to the poly(A) tails of unspliced pre-mRNAs and recruits the exosome for degradation (7,8). Studies in yeast further argue for a presumably distinct nuclear quality control mechanism that slows down or prevents release of aberrantly processed RNAs from the transcription site. Retention of transcripts at the transcription site occurs in cells with mutations in components of the 3′-end processing machinery and in factors required for mRNA export (reviewed in 9). Retention at the transcription site was also observed in mammalian cells for mutant β-globin pre-mRNAs that failed to be efficiently spliced or 3′-end processed (10), and accumulation of mutant β-globin transcripts in association with chromatin was found to be exosome-dependent (11). Co-transcriptional destruction of aberrantly processed RNAs by the exonuclease Xrn2 has also been recently reported in human cells (12).

Whether co-transcriptional RNA quality control plays a physiological role in the context of human genetic diseases is unknown. To start addressing this issue, we used patient-derived lymphoblastoid cell lines to analyze the life cycle of RNAs produced from genes that contain disease-causing splicing mutations (SM). For a subset of the mutant genes, we found lower steady-state levels of RNAs associated with chromatin and reduced transcriptional activity. Treatment of cells with cycloheximide did not alter the levels of chromatin-associated mutant RNAs, suggesting a quality control mechanism independent from NMD.

Results

Epstein–Barr virus-immortalized lymphoblastoid cell lines were used as a model to study the impact of disease-causing SM on the biogenesis of mRNA. To obtain quantitative information on RNA levels, we used a biochemical fractionation approach that allows dynamic changes in transcription or nuclear RNA degradation to be distinguished from changes in cytoplasmic steady-state mRNA levels (Fig. 1A). We optimized for lymphoblastoid cells a fractionation technique that was initially described by Wuarin and Schibler (13) and subsequently modified in the Proudfoot and Black laboratories (14,15). The protocol takes advantage of the fact that once RNA polymerase II (RNAPII) initiates transcription it forms a tight complex with the DNA template that resists treatment with urea and mild detergent. The extraction procedure does not dissociate histones from DNA and therefore the chromatin remains highly compacted and can be sedimented with associated nascent transcripts by low-speed centrifugation. Transcripts detected in the nucleoplasmic supernatant fraction are assumed to have been released from the DNA template. The efficiency of the fractionation protocol was assessed by western blotting (Fig. 1B) and RT-PCR (Fig. 1C). For the western blotting (WB) assay antibodies against lamin A/C, β-actin, U2B″ and histone H3 proteins were used (Fig. 1B). Actin was found predominantly in the cytoplasmic fraction, whereas U2 snRNP specific protein B″ (U2B″), lamin A/C and histone H3 were detected exclusively in nuclear fractions. The nucleoplasmic fraction should contain nuclear proteins that either do not associate with chromatin or are loosely attached to chromatin.

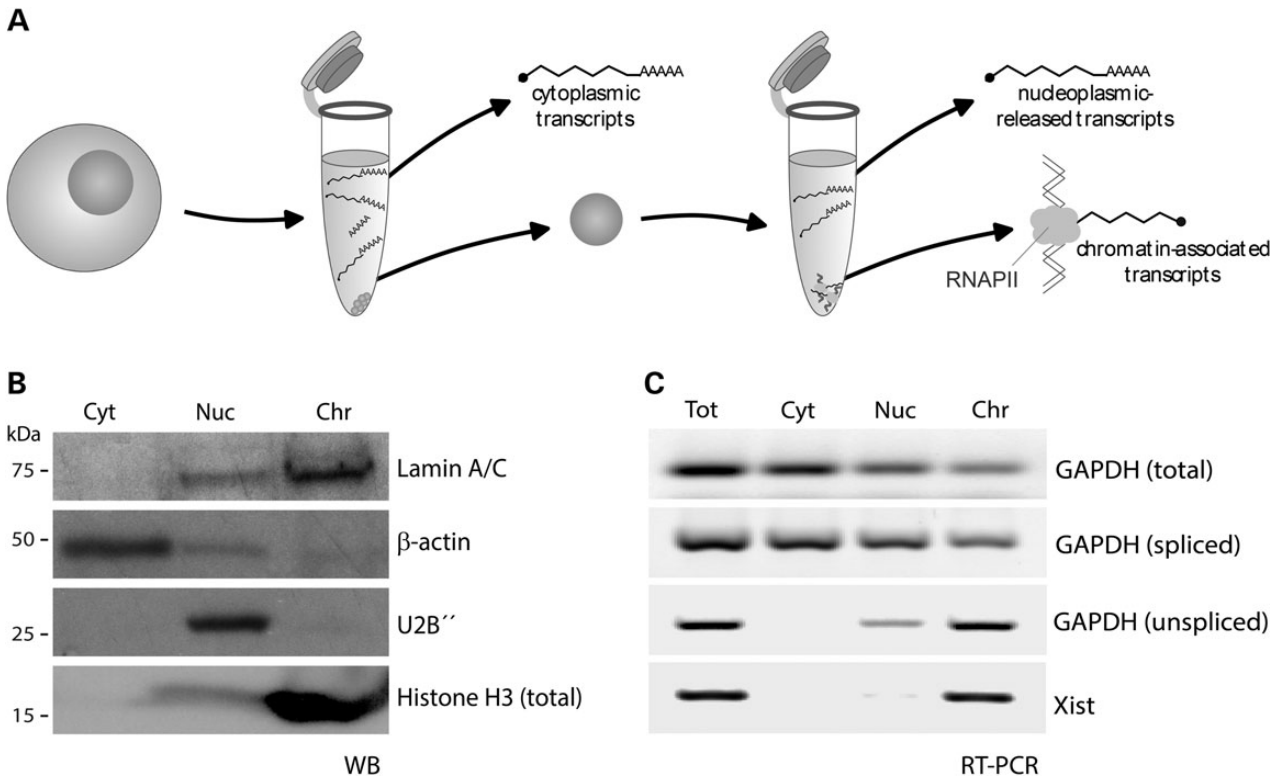


Figure 1. Sub-cellular fractionation. (**A**) Illustration of the sub-cellular fractionation procedure. After cell lysis, nuclei are separated from the cytoplasmic (cyt) fraction by centrifugation. Nuclei are then treated with urea and non-ionic detergent. Upon centrifugation, the chromatin-associated fraction (chr) sediments separating from the soluble nucleoplasmic fraction (nuc). (**B**) WB analysis. Lymphoblastoid cells (GM16113) were fractionated and analyzed by WB for detection of Lamin A/C, β-actin, U2B″ and Histone H3 (total). Equal amounts of total protein from each fraction were loaded per lane. (**C**) RT-PCR analysis. RNA was isolated from lymphoblastoid cells (GM04490), reverse transcribed with random primers and PCR amplified using primers for total, spliced and unspliced GAPDH RNA and total Xist RNA. Equal amounts of PCR product from total and fractionated samples were loaded per lane.

The U2B″ is mostly detected in the nucleoplasmic fraction as previously reported for other components of the spliceosome (14). Lamin and histone H3 are well-known chromatin-associated proteins and, accordingly, they are predominantly detected in the chromatin fraction. To characterize the RNA species present in each fraction, RT-PCR analysis was carried out with primers for total, unspliced and spliced Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) mRNA, as well as primers for Xist RNA (Fig. 1C). The results clearly show that unspliced GAPDH pre-mRNA is restricted to the nucleus and localizes predominantly in the chromatin fraction. Spliced mRNA is also detected in the chromatin fraction, consistent with the view that most splicing occurs co-transcriptionally (16), but is most abundant in the cytoplasm. The distribution of total GAPDH RNA is similar to that of spliced mRNA, as expected considering that mature transcripts are transported and accumulate in the cytoplasm. A completely different distribution pattern is observed for Xist RNA, which is restricted to the nucleus and predominantly localized in the chromatin fraction consistent with its well-established physical interaction with the X-chromosome (17).

Having validated the fractionation methodology, we next determined RNA levels in cell lines derived from a healthy donor and from patients affected by three distinct monogenic recessive disorders associated with SM: Barth syndrome (OMIM 302060), Deafness, autosomal recessive 49 (OMIM 610153) and Xeroderma Pigmentosum (OMIM 278720).

Barth syndrome is a X-linked recessive syndrome caused by mutations in the *TAZ* gene (MIM: 300394), which codes for an acyltransferase required for remodeling of cardiolipin in the inner mitochondrial membrane. TAZ loss of function results in an inborn error of lipid metabolism (18–20). We analyzed two patient-derived cell lines, each containing a point mutation that affects splicing of the *TAZ* gene. The SM localize in intron 1, at the 5′ and 3′ splice sites (Table 1 and Fig. 2A). It was previously shown that the 5′ splice site mutation activates two cryptic donor splice sites either upstream or downstream of the point mutation, and the 3′ splice site mutation can either activate a cryptic acceptor splice site within exon 2 or lead to exon 2 skipping (21). Most 5′SM transcripts expressed in lymphoblastoid cells correspond to the longer splice product, which does not disrupt the open reading frame. The less abundant shorter splice product has the open reading frame disrupted (21). The two splice products resulting from the 3′ splice site mutation are expressed at similar levels and only one has the open reading frame disrupted (21). For comparison, we analyzed a cell line with a point mutation in exon 2 that introduces a PTC without affecting splicing frame (19).

RNA levels were measured by quantitative real-time RT-PCR (qRT-PCR) using the primers indicated in Figure 2A. Figure 2B depicts the level of mutant transcripts in total cellular RNA as fold change relative to values detected using the same primer sets in cells from a healthy donor. The abundance of each PCR product was normalized to the level of GAPDH RNA detected in the same qRT-PCR run. The results show that the three mutant transcripts analyzed are significantly less abundant than wild-type (WT) *TAZ* RNA (Fig. 2B), in agreement with the loss of function phenotype observed in patients. Next we determined the levels of mutant transcripts in the cytoplasm, nucleoplasm and chromatin, using equal amounts of RNA from each fraction. The cytoplasmic levels of the three mutant *TAZ* RNAs are significantly reduced relative to the WT (Fig. 2C). However, distinct scenarios are observed in nuclear fractions (Fig. 2D and E). Mutant transcripts that contain a PTC but have normal splicing do not significantly differ from WT, suggesting that these RNAs are exclusively degraded in the cytoplasm (Fig. 2D and E, PTC). In contrast, transcripts with the 5′ splice site mutation are significantly less abundant than WT transcripts in both nucleoplasm and chromatin fractions (Fig. 2D and E, 5′SM). The level of transcripts with the 3′ splice site mutation is similar to WT in the nucleoplasm (Fig. 2D, 3′SM), but higher in the chromatin (Fig. 2E, 3′SM). This heterogeneity of results prompted us to analyze additional mutant transcripts associated with unrelated diseases.

Deafness, autosomal recessive 49 is a congenital profound sensorineural hearing loss of all frequencies, caused by dysfunction of a tricellulin protein coded by the *MARVELD2* gene (MIM: 610572). Tricellulin is a tight-junction protein that contributes to the structure and function of tricellular contacts of neighboring cells. Loss of function of this protein may selectively affect the cellular permeability to ions or small molecules, resulting in a toxic microenvironment for cochlear hair cells and subsequently ear loss (22,24). We analyzed cell lines derived from three patients, each homozygous for a distinct splice site mutation in the *MARVELD2* gene (Table 1 and Fig. 3A). The splice site mutations localize in intron 3 at the 3′ splice site, and in intron 4 at the 5′ splice site. The 5′ splice site mutations activate cryptic donor sites in intron 4, and the 3′ splice site mutation activates a cryptic acceptor site within exon 4; all the mutations lead to the production of mRNAs containing PTCs due to shifts in the open reading frame (22). For comparison, we analyzed a cell line homozygous a point mutation in exon 5 that introduces a PTC without affecting splicing (22).

Table 1. Cell lines used in this study

Cell line	Reference	Affected gene	Mutation	Transcript
GM16113		–	WT	WT
GM22129	Patient 2 (21)	TAZ, Xq28	IVS1+5G>A (5′SM)	Exon 1 cryptic, PTC Intron 1 cryptic
GM22165	Patient 4 (21)		IVS1–2A>G (3′SM)	Exon 2 cryptic, PTC Exon 2 skipped
GM22150	Patient 2 (19)		Trp79Ter (PTC)	PTC
GM20190	PKDF399 (22)	MARVELD2, 5q13.2	IVS3-1G>A (3′SM)	Exon 4 cryptic, PTC
GM20193	PKDF068 (22)		IVS4+2T>C (5′SM_1)	Intron 4 cryptics, PTC
GM20172	PKDF443 (22)		IVS4+2delTGAG (5′SM_2)	Intron 4 cryptics, PTC
GM20189	PKDF340 (22)		Arg500Ter (PTC)	PTC
GM04490	XP25BE (23)	XPC, 3p25.1	IVS11-1_IVS11-2 delAG IVS11-6_IVS11-7 InsCC (3′SM)	Intron 11 retained, PTC Exon 12 cryptic, PTC Exon 12 skipped, PTC

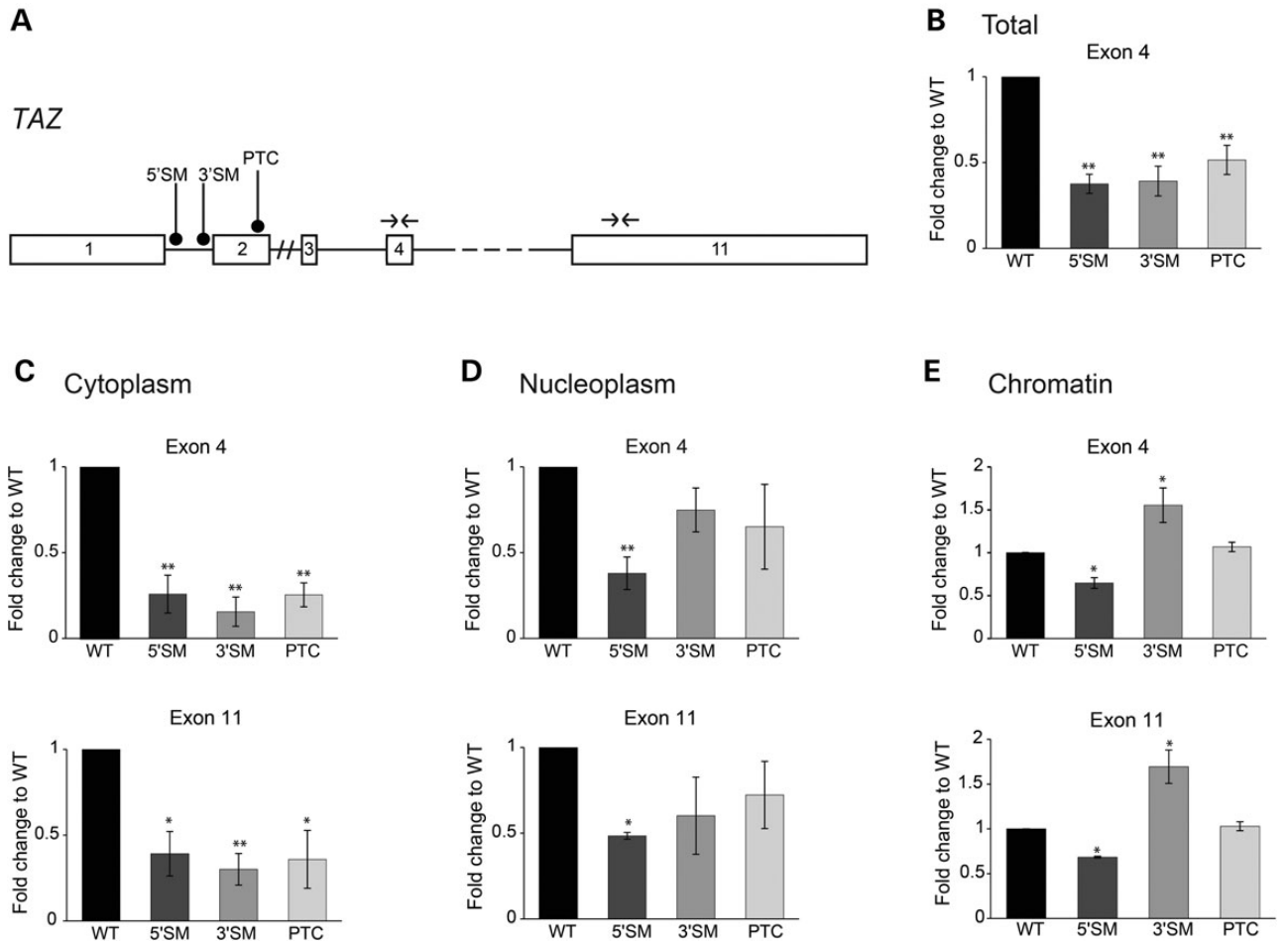


Figure 2. Sub-cellular distribution of WT and mutant *TAZ* transcripts. (**A**) Illustration of the *TAZ* gene structure (total length: 10185 bp). Exons are represented by numbered boxes and introns by lines; doubled intersected lines denote introns with more than 1000 bp. Gene region between exon 4 and exon 11 is represented by a dashed line. Positioning of mutations (5′SM, 3′SM, PTC) and primers used for PCR amplification (paired arrows) are indicated. (**B**) Total cellular RNA was extracted from the indicated cell lines, reverse transcribed with random primers and analyzed by qRT-PCR using primers for exon 4. The amount of PCR product obtained from each cell line was normalized to the level of GAPDH RNA detected in the same line. (**C–E**) RNA was extracted from sub-cellular fractions isolated from each cell line and analyzed by qRT-PCR using primer sets for exon 4 and exon 11. The amount of PCR product obtained from each fraction was normalized to the level of GAPDH RNA detected in the same fraction. In all graphs shown, data are expressed as fold change relative to the levels of WT transcripts. The histograms depict mean and standard deviation of three independent experiments. The asterisk denotes statistically significant differences (Student's *t*-test, $^{*}P < 0.05$, $^{**}P < 0.01$).

RNA levels were measured by qRT-PCR using the primers indicated in Figure 3A. Similarly to the results obtained with *TAZ* transcripts, the total cellular levels of the four mutant *MARVELD2* RNAs are significantly reduced compared with WT (Fig. 3B). Analysis of RNA levels in sub-cellular fractions reveals that mutant transcripts are significantly less abundant in the cytoplasm (Fig. 3C), in agreement with the finding that they all contain PTCs. In nuclear fractions the levels of mutant transcripts that contain a PTC but have normal splicing are similar to WT (Fig. 3D and E, PTC), indicating that these RNAs are exclusively degraded in the cytoplasm. However, all transcripts with SM are significantly less abundant in the nucleoplasm (Fig. 3D). In the chromatin fraction, significantly reduced levels are only detected for the 3′ splice site mutant (Fig. 3E, 3′SM).

As a third model we analyzed cells from a patient with Xeroderma pigmentosum, an autossomic recessive condition characterized by increased sensitivity to ultraviolet irradiation and increased risk of skin cancer. It is caused by mutations in the *XPC* gene (MIM: 613208), which encodes a protein required for DNA repair (23,25). The cell line analyzed is homozygous for two distinct mutations at the 3′ splice site of intron 11 (Table 1 and Fig. 4A). These mutations lead to skipping of exon 12, retention of intron 11 and activation of a 3′ cryptic splice site in exon 12, resulting in introduction of PTCs (23). Quantitative real-time RT-PCR using the primers indicated in Figure 4A reveals a significant reduction in the total cellular levels of mutant *XPC* RNA compared with WT (Fig. 4B). Analysis of sub-cellular fractions further shows that mutant transcripts are significantly less abundant in the cytoplasm, nucleoplasm and chromatin (Fig. 4C, D and E).

Altogether these results show that SM are consistently associated with reduced mRNA levels in the cytoplasm and, for a subset of mutations, down-regulation of expression is also detected in the nucleus. In contrast, mRNAs resulting from point mutations that introduce a PTC but do not interfere with splicing appear exclusively down-regulated in the cytoplasm. To determine whether lower steady-state RNA levels in the nucleus result from reduced transcription of genes containing SM, we measured newly transcribed RNA levels by metabolic labeling with the natural uridine derivative 4-thiouridine (4sU). This approach

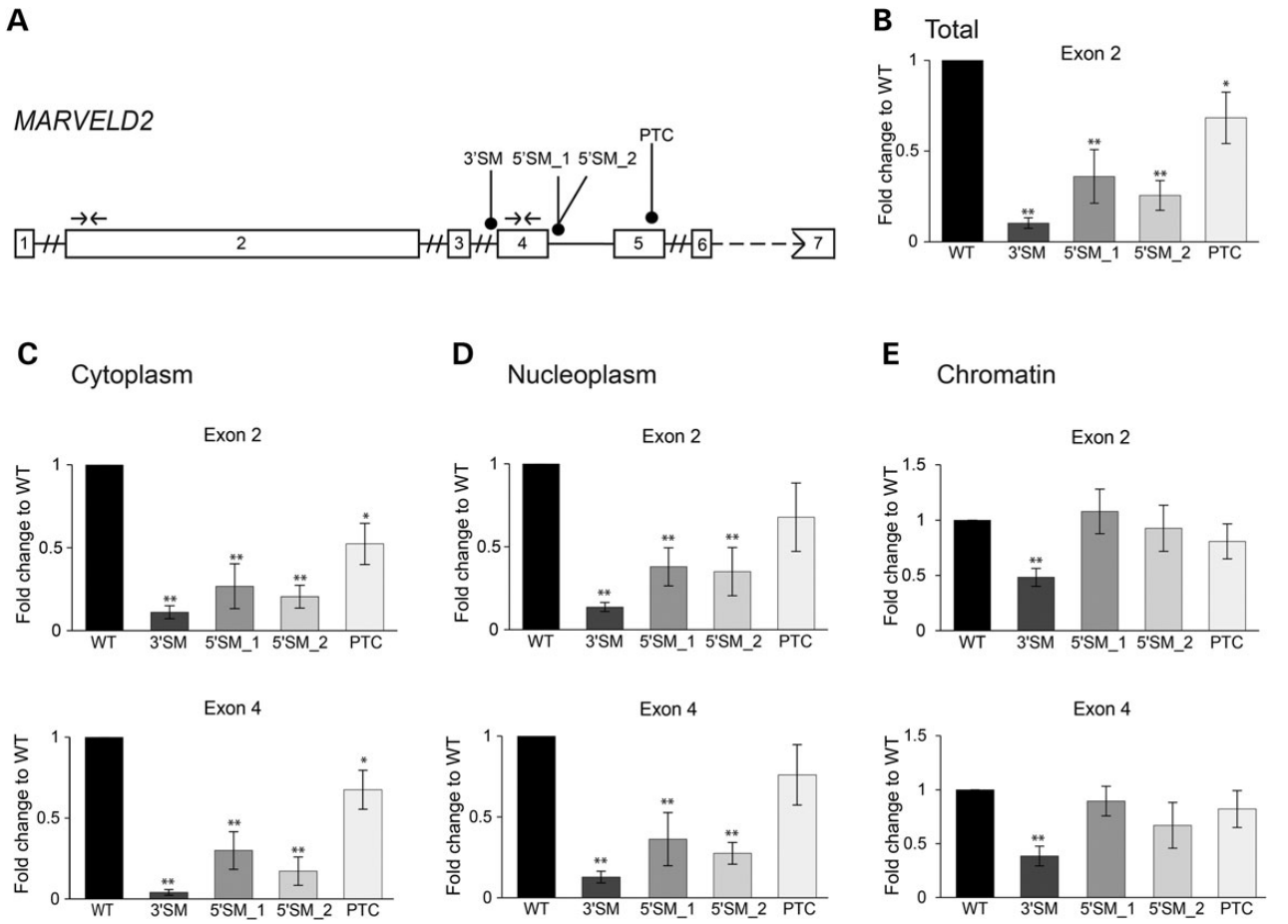


Figure 3. Sub-cellular distribution of WT and mutant *MARVELD2* transcripts. (**A**) Illustration of the *MARVELD2* gene structure (total length: 27762 bp). Exons are represented by numbered boxes and introns by lines; doubled intersected lines denote introns with more than 1000 bp. Intron 6 and part of exon 7 are represented by a dashed line. Positioning of mutations (3′SM, 5′SM_1, 5′SM_2, PTC) and primers used for PCR amplification (paired arrows) are indicated. (**B**) Total cellular RNA was extracted from the indicated cell lines, reverse transcribed with random primers and analyzed by qRT-PCR using primers for exon 2. The amount of PCR product obtained from each cell line was normalized to the level of GAPDH RNA detected in the same line. (**C–E**) RNA was extracted from sub-cellular fractions isolated from each cell line and analyzed by qRT-PCR using primer sets for exon 2 and exon 4. The amount of PCR product obtained from each fraction was normalized to the level of GAPDH RNA detected in the same fraction. In all graphs shown, data are expressed as fold change relative to the levels of WT transcripts. The histograms depict mean and standard deviation of three independent experiments. The asterisk denotes statistically significant differences (Student's *t*-test, $^{*}P < 0.05$, $^{**}P < 0.01$).

provides direct access to newly synthesized transcripts with minimal toxic effects (26), although it may induce a nucleolar stress response (27). Nascent RNA was labeled by adding 4sU to the cell culture medium for 10 min followed by isolation of total cellular RNA. Newly transcribed RNA species containing thiol-groups were then biotinylated, purified using streptavidin-coated beads and analyzed by qRT-PCR (Fig. 5A). As RNAPII transcribes with elongation rates ranging between 0.5 and 4 kb/min (29), synthesis of new *TAZ* RNAs may take from 2.5 to 20 min, whereas *MARVELD2* and *XPC* RNAs may require between 7 or 8 min to approximately 1 h. Thus, we expect that after incubation with 4sU for 10 min, most labeled RNAs are in the process of being synthesized and therefore should be confined to the chromatin fraction. The results shown in Figure 5B are in very good agreement with this prediction. To assess the extent to which transcription of the *TAZ*, *MARVELD2* and *XPC* genes differs between lymphoblastoid cell lines derived from normal individuals, we analyzed a recently reported microarray dataset (28). The results show that the transcription rate of these genes is similar across cells from three distinct individuals (Fig. 5C). Next, we compared the levels of nascent transcripts produced by WT and mutant genes using primers to amplify both exonic and intronic regions of *TAZ* (Fig. 5D), *MARVELD2* (Fig. 5E) and *XPC* (Fig. 5F) transcripts. A significant down-regulation of nascent transcripts is observed for the *TAZ* 5′ splice site (Fig. 5D, 5′SM) and *MARVELD2* 3′ splice site (Fig. 5E, 3′SM) mutants, strongly suggesting that these genes are less efficiently transcribed. No evidence for reduced transcriptional activity of the *XPC* 3′ splice site mutant gene is observed, arguing that the lower steady-state RNA levels detected in the chromatin fraction likely reflect rapid nuclear degradation of these transcripts.

To determine the contribution of NMD to the observed down-regulation of mutant RNAs in each sub-cellular fraction, cells were treated with cycloheximide (CHX), a drug that inhibits translation and hence indirectly blocks NMD (30). After 3 h of treatment, cells were fractionated and changes in RNA levels analyzed by qRT-PCR. RNA levels in each treated fraction (CHX+) are expressed as fold change relative to the levels in the corresponding non-treated fraction (CHX-; Figs 6–8A). Alternatively, mutant RNA levels in each treated fraction (CHX+) are expressed as fold change relative to the levels of WT transcripts in the corresponding fraction from treated cells (Figs 6–8B). Analysis

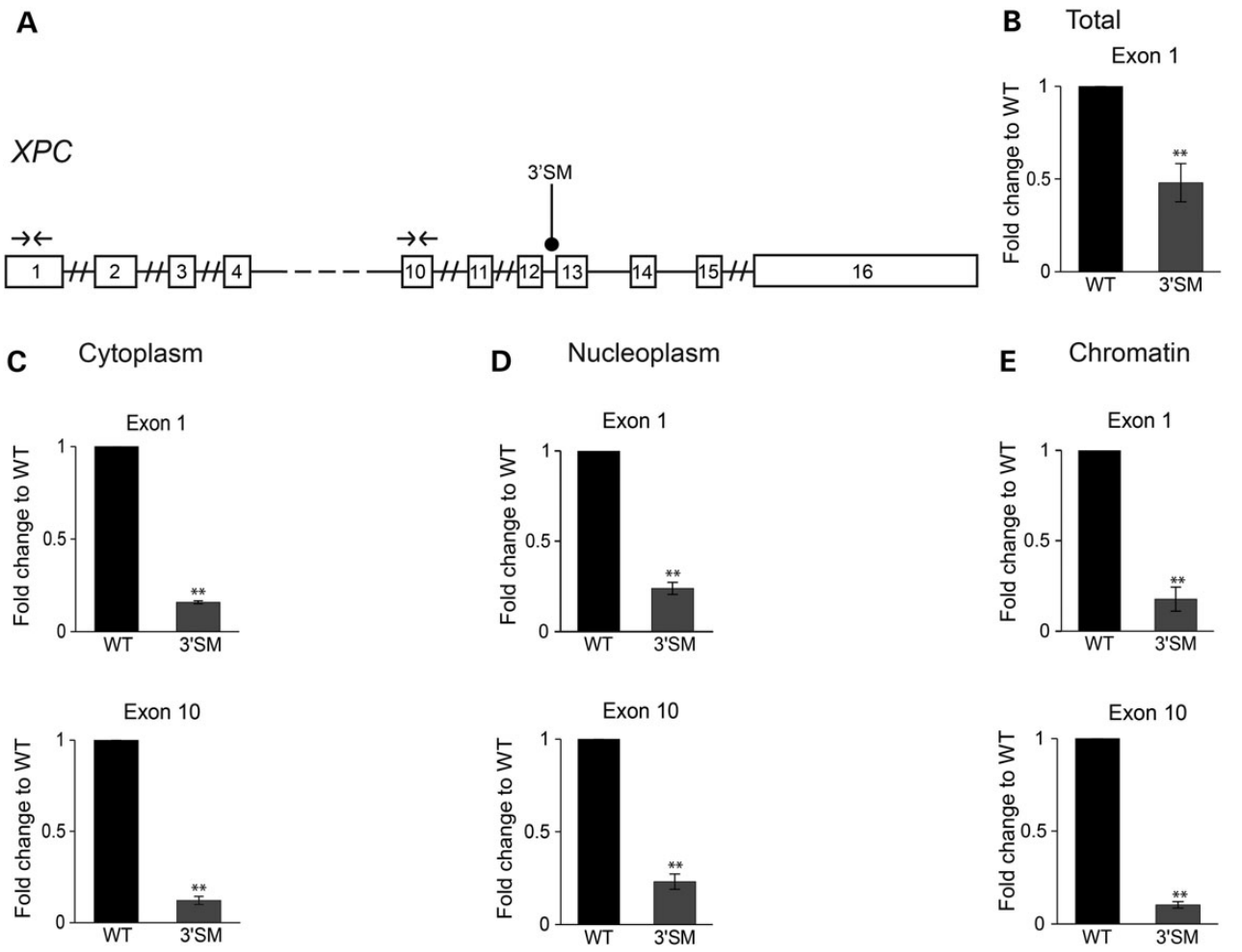


Figure 4. Sub-cellular distribution of WT and mutant *XPC* transcripts. (**A**) Illustration of the *XPC* gene structure (total length: 33525 bp). Exons are represented by numbered boxes and introns by lines; doubled intersected lines denote introns with more than 1000 bp. Gene region between exon 4 and exon 10 is represented by a dashed line. Positioning of the mutation (3′SM) and primers used for PCR amplification (paired arrows) are indicated. (**B**) Total cellular RNA was extracted from the indicated cell lines, reverse transcribed with random primers and analyzed by qRT-PCR using primers for exon 1. The amount of PCR product obtained from each cell line was normalized to the level of GAPDH RNA detected in the same line. (**C–E**) RNA was extracted from sub-cellular fractions isolated from each cell line and analyzed by qRT-PCR amplified using primer sets for exon 1 and exon 10. The amount of PCR product obtained from each fraction was normalized to the level of GAPDH RNA detected in the same fraction. In all graphs shown, data are expressed as fold change relative to the levels of WT transcripts. The histograms depict mean and standard deviation of three independent experiments. The asterisk denotes statistically significant differences (Student's *t*-test, $^{**}P < 0.01$).

of *TAZ* (Fig. 6), *MARVELD2* (Fig. 7) and *XPC* (Fig. 8) mutant and WT transcripts shows that treatment with CHX consistently results in an increase in RNA levels in the cytoplasm. This increase is most obvious for mutant transcripts, as expected since their degradation by NMD is most probably impaired by CHX. An exception is the *MARVELD2* PTC mutant, which gives rise to RNAs that are not affected by CHX, suggesting that they escape NMD. Accordingly, this particular mutant has been described to encode a truncated tricellulin protein (22). The finding that CHX induces accumulation of WT transcripts is also in agreement with previous reports (31,32).

An accumulation of both WT and mutant RNAs is further detected in the nucleoplasm of CHX treated cells. This observation argues that the lower steady-state levels of mutant transcripts observed in association with the nucleoplasm without a corresponding decrease in the chromatin fraction could be due to contamination of the nucleoplasmic fraction by mRNAs that have already been exported from the nucleus but remain associated with the cytoplasmic side of the nuclear envelope, as previously proposed (4). In contrast, CHX does not significantly alter the levels of WT and mutant RNAs associated with the chromatin fraction. However, the levels of *TAZ* 5′SM, *MARVELD2* 3′SM and *XPC* 3′SM RNAs persist reduced compared with WT in the chromatin fraction of CHX treated cells (Figs 6–8B). Noteworthy, *TAZ* 5′SM and *MARVELD2* 3′SM RNAs, which are less efficiently transcribed (Fig. 5D and E), respond less to CHX treatment than other mutant forms of the same gene. The mild effect of CHX on cytoplasmic levels of *TAZ* 5′SM transcripts is in agreement with the finding that the majority of these RNAs are devoid of PTCs and therefore should not be degraded by NMD. Taken together, these observations suggest that some SM result in RNAs that are primarily degraded by NMD in the cytoplasm, while others can be targeted by transcription-coupled quality control mechanisms that operate independently from NMD.

Discussion

The results reported in this study suggest that splice site mutations in human cells trigger chromatin-associated RNA surveillance responses that contribute to down-regulate the

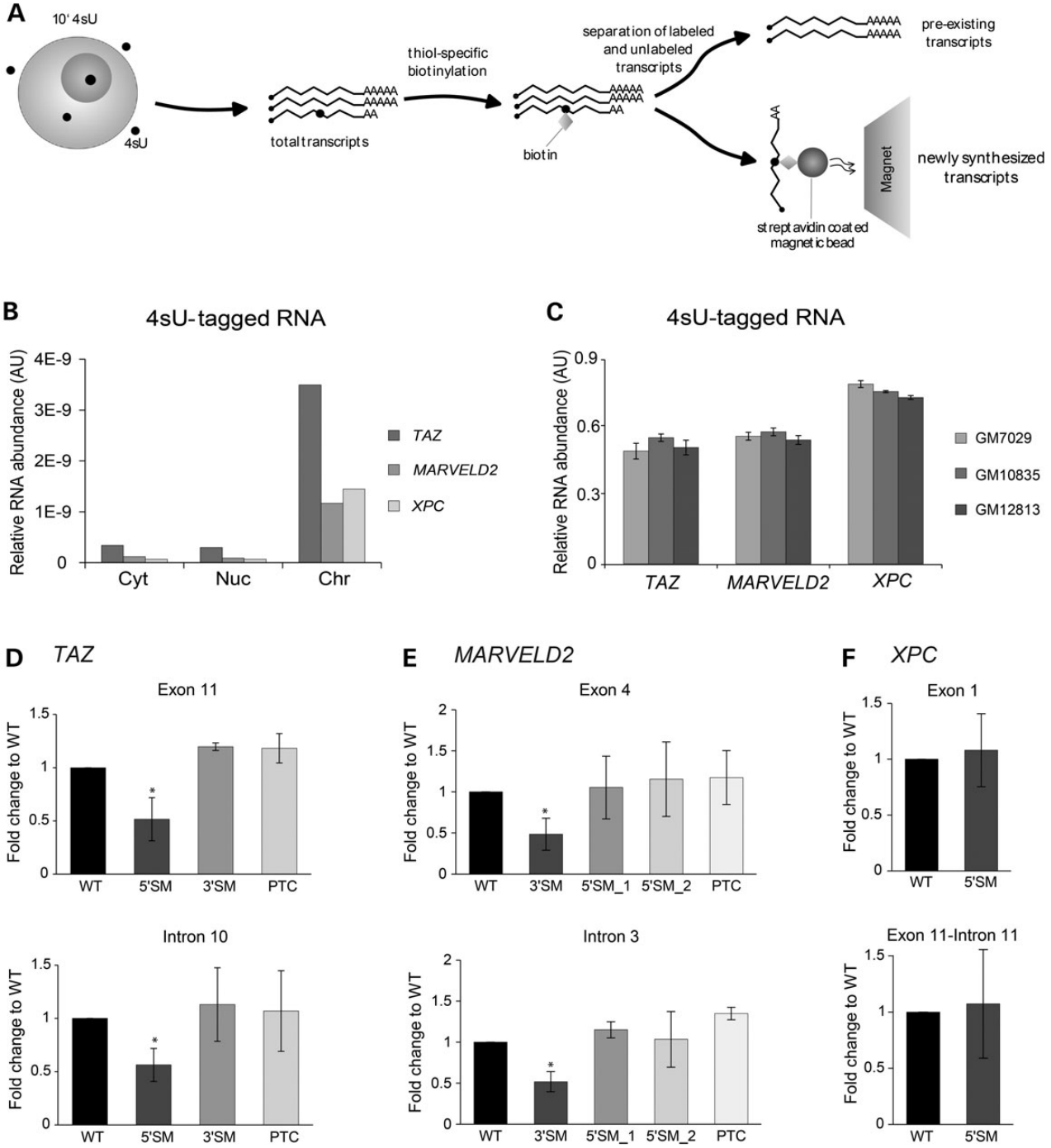


Figure 5. Analysis of nascent RNA by metabolic labeling. (**A**) Illustration of the metabolic labeling procedure. Cells in culture are incubated with 4-thiouridine (4sU). Total cellular RNA is extracted and thiol-containing molecules are biotinylated. Biotinylated RNA is then purified using streptavidin-coated magnetic beads. (**B**) Sub-cellular localization of 4sU-tagged RNA. Cells from a healthy donor (WT) were incubated with 4sU for 10 min and fractionated (Cyt: cytoplasm; Nuc: nucleoplasm; Chr: chromatin). RNA tagged with 4sU was purified from each fraction and analyzed by qRT-PCR as described in figures 2, 3 and 4. AU (arbitrary units). (**C**) Inter-individual differences of 4sU-tagged RNA. Nascent RNAs were isolated from lymphoblastoid cell lines derived from three unrelated healthy individuals (GM7029, GM10835 and GM12813) after incubation with 4sU for 2 h (analysis of GSE34204 dataset, (28)). The amount of labeled *TAZ*, *MARVELD2* and *XPC* RNA was normalized to the level of labeled *GAPDH* RNA detected in the same cell line. The histogram depicts mean and standard deviation of three biological replicates (independent cell cultures). AU (arbitrary units). (**D–F**) Quantification of nascent transcripts produced by WT and mutant genes. Cells were incubated with 4sU for 10 min. Total 4sU-tagged RNA was purified and analyzed by qRT-PCR using primers that recognize exonic (top) or intronic (bottom) regions. The amount of PCR product in each cell type was normalized to the level of *GAPDH* RNA detected in the same cell type. Data are expressed as fold change relative to the levels of WT transcripts. The histograms depict mean and standard deviation of three independent experiments. The asterisk denotes statistically significant differences (Student's *t*-test, $^*P < 0.05$).

expression of abnormal mRNAs independently of NMD. We analyzed six cell lines derived from patients carrying SM and in three of them we found reduced mutant RNA levels associated with chromatin. In two of these lines, lower abundance of mutant chromatin-associated RNA correlated with reduced transcriptional activity.

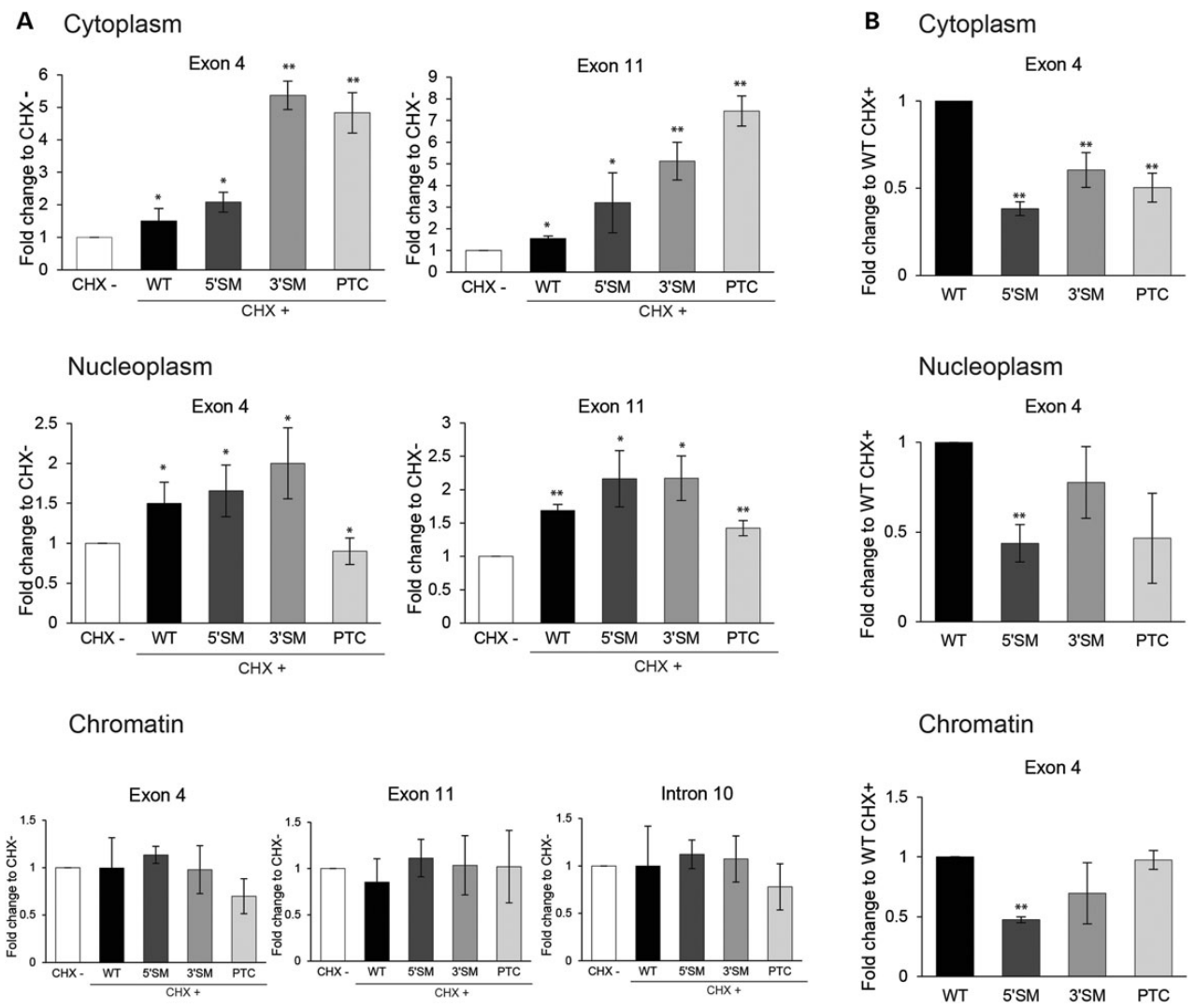


Figure 6. Effect of cycloheximide on *TAZ* transcripts. Cells were either non-treated (CHX-) or treated with cycloheximide for 3 h (CHX+). The levels of WT and mutant transcripts in each sub-cellular fraction were analysed by qRT-PCR using the indicated primer sets. The amount of PCR product was always normalized to the level of *GAPDH* RNA. Data are expressed as fold change relative to the levels of non-treated cells (**A**) or as fold change relative to the levels of WT transcripts in treated cells (**B**). The histograms depict mean and standard deviation of three independent experiments. The asterisk denotes statistically significant differences (Student's *t*-test, $^{*}P < 0.05$, $^{**}P < 0.01$).

A link between SM and transcription was previously described (33,34). In the study by Damgaard *et al.*, mutations in the promoter-proximal 5′ splice site were shown to severely decrease transcription by a mechanism that involved U1 snRNA recognition and assembly of the preinitiation complex (33). Here we observe a similar scenario for the *TAZ* 5′ splice site mutant. However, we also detected decreased transcription of the *MARVELD2* 3′SM gene, which contains a 3′ splice site mutation in the third intron. This observation raises the possibility that additional mechanisms are involved in coupling transcription to splicing efficiency. Indeed, inefficient splicing can cause stalling of spliceosomes on the transcripts, leading to recruitment of the RNAi machinery, heterochromatin formation and transcriptional silencing (35,36). Down-regulating the transcription of mutant genes appears 'economical', as it saves energy in producing and discarding aberrant RNAs. Yet, many transcripts produced from genes with SM escape this type of control.

Although transcription from the *TAZ* 3′SM and *XPC* 3′SM genes is similar to WT, RNA levels associated with chromatin differ significantly. The steady-state level of chromatin-associated *TAZ* 3′SM transcripts is higher than WT, whereas *XPC* 3′ SM transcripts are reduced compared with WT. The results obtained with *TAZ* 3′SM transcripts are reminiscent of our previous observations with β-globin splicing mutants (10,11), suggesting that abnormally processed RNAs persist associated with the chromatin template and consequently accumulate in this fraction. In contrast, the results obtained with *XPC* 3′SM suggest that these transcripts undergo a fast co-transcriptional decay most likely mediated by the exosome and/or Xrn2 (12).

A main conclusion from this study is that disease-causing SM can have a variety of effects on mRNA biogenesis. For all disease-associated genes analyzed, a single splice site mutation leads to expression of multiple mRNA isoforms. Some of these isoforms may contain a PTC due to a frame shift caused by activation of a cryptic splice site or exon skipping, others may be recognized as abnormally spliced due to intron retention, while others may not be recognized as faulty (namely, if the reading frame is not disrupted). Thus, depending on the isoform expressed, the

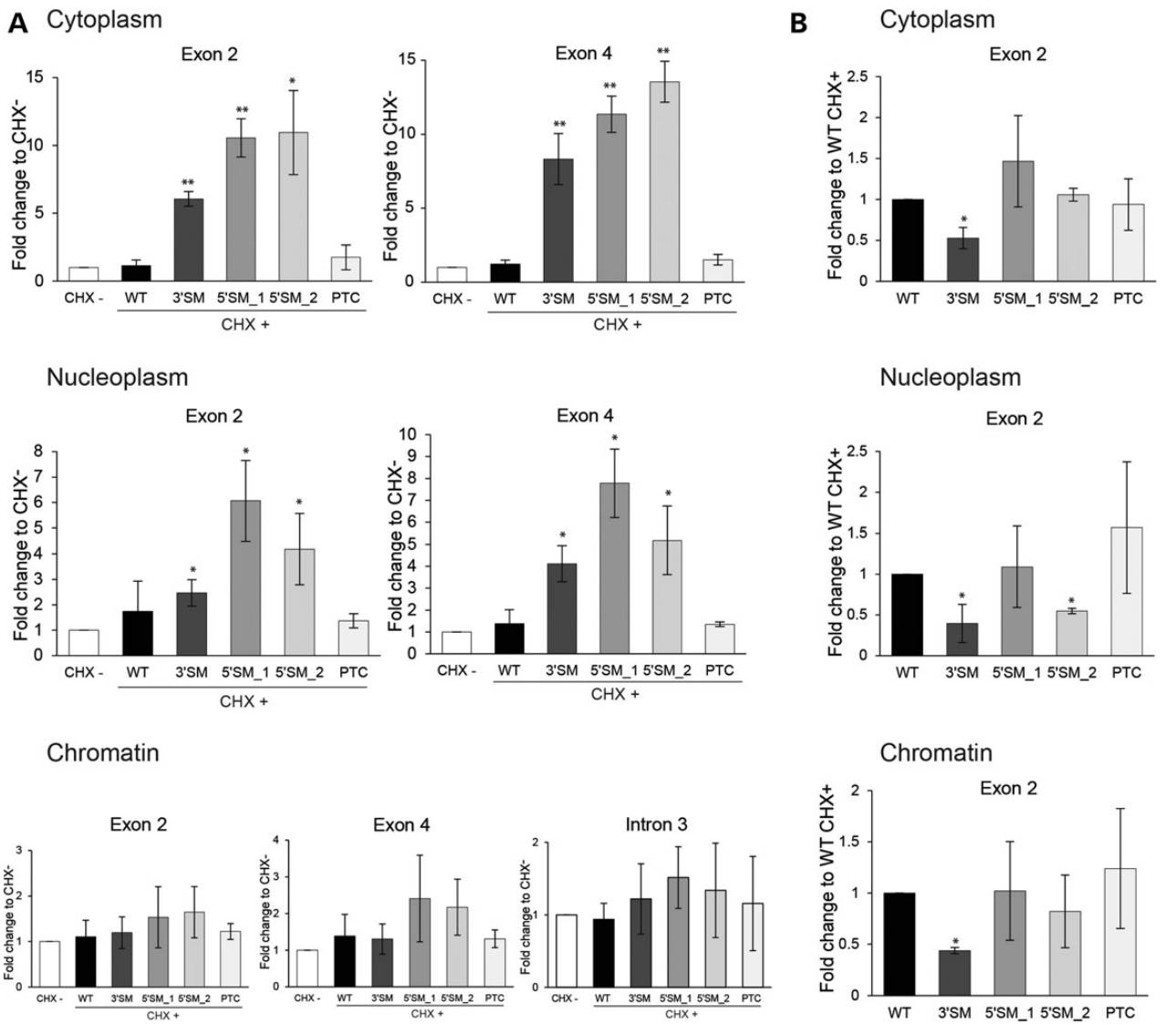


Figure 7. Effect of cycloheximide on *MARVELD2* transcripts. Cells were either non-treated (CHX-) or treated with cycloheximide for 3 h (CHX+). The levels of WT and mutant transcripts in each sub-cellular fraction were analysed by qRT-PCR using the indicated primer sets. The amount of PCR product was always normalized to the level of *GAPDH* RNA. Data are expressed as fold change relative to the levels of non-treated cells (**A**) or as fold change relative to the levels of WT transcripts in treated cells (**B**). The histograms depict mean and standard deviation of three independent experiments. The asterisk denotes statistically significant differences (Student's *t*-test, $^{*}P < 0.05$, $^{**}P < 0.01$).

mutant RNAs may be differentially recognized by distinct surveillance mechanisms. We also found that *TAZ*, *MARVELD2* and *XPC* genes are expressed at low levels in immortalized lymphoblastoid cells. Since the proteins encoded by these genes have tissue-specific functions, it remains to be established whether the patterns of mRNA biogenesis observed in lymphoblastoid cells are physiologically representative. Another limitation of working with immortalized lymphoblastoid cell lines is that these cells were resistant to RNA interference manipulations aimed at identifying the nucleases responsible for mutant RNA degradation in the nucleus. For future studies, induced pluripotent stem cells (iPSCs) derived from patients are likely to represent improved disease models. Differentiation of iPSCs into the specific cell types that require expression of the mutant genes for their normal function will provide a valuable system to address how cytoplasmic and nuclear quality control mechanisms operate to reduce expression of abnormal RNAs caused by SM.

In summary, our data supports the view that multiple layers of surveillance occur both in the nucleus and in the cytoplasm to minimize potentially toxic effects caused by faulty mRNAs. Although it is not yet possible to predict which SM will target RNAs for co-transcriptional surveillance, we expect this work will contribute to open new research venues addressing the impact of transcription-coupled non-NMD quality control pathways in the context of human genetic diseases. Ultimately, understanding how disease-causing SM are recognized by cellular quality control mechanisms may help in the rational design of more effective therapies for these disorders.

Materials and Methods

Cells and drug treatment

Lymphoblastoid cell lines immortalized by Epstein–Barr virus infection were obtained from the NIGMS Human Genetic Cell

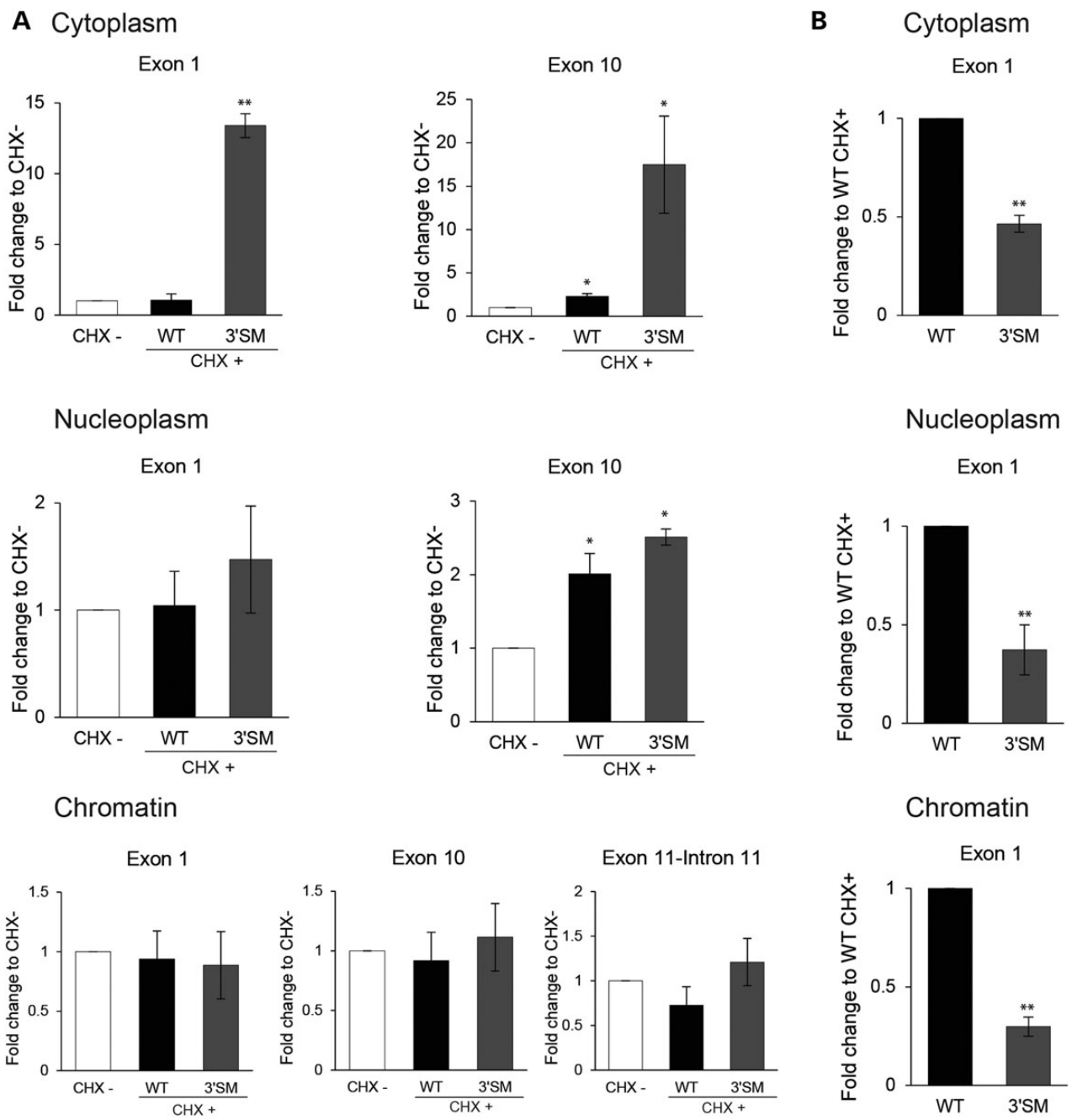


Figure 8. Effect of cycloheximide on *XPC* transcripts. Cells were either non-treated (CHX-) or treated with cycloheximide for 3 h (CHX+). The levels of WT and mutant transcripts in each sub-cellular fraction were analysed by qRT-PCR using the indicated primer sets. The amount of PCR product was always normalized to the level of *GAPDH* RNA. Data are expressed as fold change relative to the levels of non-treated cells (**A**) or as fold change relative to the levels of WT transcripts in treated cells (**B**). The histograms depict mean and standard deviation of three independent experiments. The asterisk denotes statistically significant differences (Student's *t*-test, $^{*}P < 0.05$, $^{**}P < 0.01$).

Repository collections of the Coriell Institute for Medical Research, USA. Barth syndrome cell lines are GM22129; GM22165; and GM22150. Deafness, autosomal recessive 49 cell lines are GM20190; GM20193; GM20172; GM20189 and Xeroderma Pigmentosum cell line is GM04490. The healthy donor cell line is GM16113. The cell lines are described in detail in Table 1. Cells were cultured in RPMI 1640 medium supplemented with 18% heat-inactivated serum, 2 mM non-essential amino acid solution and 2 mM L-Glutamin at 37°C in 5% CO_2. All cell culture reagents were from Gibco, UK. Cells were treated with 50 µg/ml cycloheximide (C7698, Sigma, USA) for 3 h at 37°C.

Total RNA isolation and sub-cellular fractionation

Nuclear and cytoplasmic RNA fractions were isolated as described (37). Briefly, cells were incubated in RSB buffer (10 mM Tris, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$) for swelling, centrifuged and resuspended in RSBG40 buffer (10 mM Tris, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 10% glycerol, 0.2% Nonidet P-40, 0.5 mM dithiothreitol and 40 U/ml RNase) for lyses of the cell membrane. The fractionation of the nuclei into chromatin-associated and nucleoplasmic RNA was adapted from (13–15). The nuclear pellet was gently resuspended in a prechilled glycerol buffer (20 mM

Tris pH 7.9, 75 mM NaCl, 0.5 mM EDTA, 0.85 mM DTT, 0.125 mM PMSF and 50% glycerol) and an equal volume of cold nuclei lysis buffer (10 mM HEPES pH7.6, 300 mM NaCl, 0.2 mM EDTA, 1 mM DTT, 7.5 mM $MgCl_2$, 1 M Urea and 1% NP-40) was added. The tube was gently vortexed for 2 × 2 s and incubated for 10 min on ice. Chromatin was pelleted and incubated in 10 mM Tris pH 7.5, 500 mM NaCl, 10 mM $MgCl_2$, 100 U/µl DNase I, 100 U/µl RNase OUT. RNA was extracted from each fraction and from the whole cell using PureZOL RNA isolation reagent (Bio-Rad, USA).

Immunoblotting

Immunoblotting of proteins extracted from each sub-cellular fraction was previously described (11). The following primary antibodies were used: anti-lamin A/C (H-110, Santa Cruz Biotechnology, Inc); anti β-actin (Sigma); anti-U2B″ (clone 4G3, PROGEN Biotechnik GmbH); and anti-histone H3 (Abcam).

4sU Labeling

Nascent RNA was labeled with 4-thiouridine (Sigma, USA) as described (38). Total RNA was extracted using PureZOL (Bio-Rad, USA). Thiol-labeled RNA was biotinylated using EZ-Link Biotin-HPDP (Pierce, USA) and separated from untagged species using µMACS streptavidin-coated magnetic beads and columns (Miltenyi, Germany).

Quantitative real-time PCR

RNA was used as template for cDNA synthesis using random primers from the High Fidelity Kit (Roche) according to the manufacturer's instructions. PCR reactions were performed in the ViiA™ 7 Real-Time PCR System (Applied Biosystems, USA), using iTaq Universal SYBR Green Supermix (Bio-Rad, USA). Gene-specific primers are presented in Table 2. Each sample was run in duplicate. The $2^{-\Delta Ct}$ method (39) was used to measure the relative changes in transcript levels.

Table 2. Sequence of primers used in PCR experiments

Primer name	Sequence
qRT-PCR primers	
GAPDH For	GAAGGTGGAGGTCGGAGTC
GAPDH Rev	GAAGATGGTGATGGGATTTC
TAZ ex.4 For	AGACATCTGCTTCACCAAGGAGCTA
TAZ ex.4 Rev	TCGGCACACAGGCACACACT
TAZ ex. 11 For	TGCGGAAAGCCCTGACGGA
TAZ ex. 11 Rev	GGCTGGAGGTGGTTGTGGAGC
TAZ int.10 For	GCCTCCACCCTCTCCATCCCG
TAZ int.10 Rev	TGCACCCCTCGGGAAGCTTGG
MARVELD2 ex.2 For	CTCCAGCAAGACCAAACCAC
MARVELD2 ex.2 Rev	CAGCCTCTTTCCGGGAACTA
MARVELD2 ex.4 For	GGTGACAGACAAAGAGACTCAG
MARVELD2 ex.4 Rev	ACATAGTCGGGCATCACGAT
MARVELD2 int.3 For	AGGTGATCTGGCTTCTGTCC
MARVELD2 int.3 Rev	TGGATTAGGTGTGGAGGCTG
XPC ex. 1 For	GGCCGGCGTTCTAGCGCAT
XPC ex. 1 Rev	CACGCCGGGCCTTGCTCTTG
XPC ex. 10 For	GGCTAAACACATGGACCAGC
XPC ex. 10 Rev	GTAGACCGCTTCTCCACGAC
XPC ex.11_int.11	AGGCTTGGAGAAGTACCCTACAAG
XPC ex.11_int.11	TGAATCCTGCTCAAGCCGGGAAA
RT-PCR primers	
GAPDH ex.3 For	TCACCAGGGCTGCTTTTAAC
GAPDH ex.3 Rev	CATGTAGTTGAGGTCAATGAAGG
GAPDH ex.5 Rev	TGAAGACGCCAGTGGAC
GAPDH int.2 For	GGGAAGGAAATGAATGGGCAG
GAPDH int.2 Rev	GGACCTCCATAAACCCACTT
Xist For	GTCAGGAGAAAGAAGTGGAGGG
Xist Rev	ACAGAGGAATGGAGGGAGGTT

Microarray data analysis

Data deposited in GEO database with the reference GSE34204 (28), were used for analysis. Microarrays were processed by using the AltAnalyze software version 2.0.8 (40). Briefly, raw CEL data files from the deposited microarrays were normalized by the RMA algorithm. Probesets with detection above background (DABG) p-values above 0.5 or non-logarithmic expression below 1.0 were removed from the analysis. Gene expression levels were determined using only constitutive probesets, using the gene annotation present in AltAnalyze derived from Ensembl (41) and USCS (42) databases.

Acknowledgements

We thank Sérgio Marinho, Marisa Cabrita, Ana Jesus and Dinora Levy for excellent technical help, and our colleagues Teresa Carvalho and Sérgio de Almeida for reading the manuscript. We also want to thank Célia Carvalho, Joana Desterro, Catarina Santos, Catarina Alves do Vale, Rita Almeida, Tomás Gomes, Ana Pena and Vasco Neves for insightful discussions and support.

Conflict of Interest statement: The authors declare that there are no conflict of interests.

Funding

This work was supported by Fundação para a Ciência e a Tecnologia (grant PTDC/BIA-BCM/101575/2008 and fellowship SFRH/BD/90231/2012 to R.V.D.).

References

1. Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A. and Cooper, D.N. (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.*, **133**, 1–9.
2. Lim, K.H., Ferraris, L., Filloux, M.E., Raphael, B.J. and Fairbrother, W.G. (2011) Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc. Natl. Acad. Sci. U S A*, **108**, 11093–11098.
3. Singh, R.K. and Cooper, T.A. (2012) Pre-mRNA splicing in disease and therapeutics. *Trends Mol. Med.*, **18**, 472–482.
4. Popp, M.W. and Maquat, L.E. (2013) Organizing principles of mammalian nonsense-mediated mRNA decay. *Annu. Rev. Genet.*, **47**, 139–165.
5. Kilchert, C. and Vasiljeva, L. (2013) mRNA quality control goes transcriptional. *Biochem. Soc. Trans.*, **41**, 1666–1672.
6. Bousquet-Antonelli, C., Presutti, C. and Tollervey, D. (2000) Identification of a regulated pathway for nuclear pre-mRNA turnover. *Cell*, **102**, 765–775.
7. Schmid, M., Poulsen, M.B., Olszewski, P., Pelechano, V., Saguez, C., Gupta, I., Steinmetz, L.M., Moore, C. and Jensen, T.H. (2012) Rrp6p controls mRNA poly(A) tail length and its decoration with poly(A) binding proteins. *Mol. Cell.*, **47**, 267–280.

8. Lemieux, C., Marguerat, S., Lafontaine, J., Barbezier, N., Bahler, J. and Bachand, F. (2011) A Pre-mRNA degradation pathway that selectively targets intron-containing genes requires the nuclear poly(A)-binding protein. *Mol. Cell.*, **44**, 108–119.
9. Porrua, O. and Libri, D. (2013) RNA quality control in the nucleus: the Angels' share of RNA. *Biochim. Biophys. Acta*, **1829**, 604–611.
10. Custodio, N., Carmo-Fonseca, M., Geraghty, F., Pereira, H.S., Grosveld, F. and Antoniou, M. (1999) Inefficient processing impairs release of RNA from the site of transcription. *EMBO J.*, **18**, 2855–2866.
11. de Almeida, S.F., Garcia-Sacristan, A., Custodio, N. and Carmo-Fonseca, M. (2010) A link between nuclear RNA surveillance, the human exosome and RNA polymerase II transcriptional termination. *Nucleic Acids Res.*, **38**, 8015–8026.
12. Davidson, L., Kerr, A. and West, S. (2012) Co-transcriptional degradation of aberrant pre-mRNA by Xrn2. *EMBO J.*, **31**, 2566–2578.
13. Wuarin, J. and Schibler, U. (1994) Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. *Mol. Cell. Biol.*, **14**, 7219–7225.
14. Pandya-Jones, A. and Black, D.L. (2009) Co-transcriptional splicing of constitutive and alternative exons. *RNA*, **15**, 1896–1908.
15. Dye, M.J., Gromak, N. and Proudfoot, N.J. (2006) Exon tethering in transcription by RNA polymerase II. *Mol. Cell.*, **21**, 849–859.
16. Bentley, D.L. (2014) Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.*, **15**, 163–175.
17. Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S. *et al.* (2013) The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science*, **341**, 1237973.
18. Bione, S., D'Adamo, P., Maestrini, E., Gedeon, A.K., Bolhuis, P.A. and Toniolo, D. (1996) A novel X-linked gene, G4.5. is responsible for Barth syndrome. *Nat. Genet.*, **12**, 385–389.
19. Gonzalez, I.L. (2005) Barth syndrome: TAZ gene mutations, mRNAs, and evolution. *Am. J. Med. Genet. A*, **134**, 409–414.
20. Kirwin, S.M., Manolakos, A., Barnett, S.S. and Gonzalez, I.L. (2014) Tafazzin splice variants and mutations in Barth syndrome. *Mol. Genet. Metab.*, **111**, 26–32.
21. Johnston, J., Kelley, R.I., Feigenbaum, A., Cox, G.F., Iyer, G.S., Funanage, V.L. and Proujansky, R. (1997) Mutation characterization and genotype-phenotype correlation in Barth syndrome. *Am. J. Hum. Genet.*, **61**, 1053–1058.
22. Riazuddin, S., Ahmed, Z.M., Fanning, A.S., Lagziel, A., Kitajiri, S., Ramzan, K., Khan, S.N., Chattaraj, P., Friedman, P.L., Anderson, J.M. *et al.* (2006) Tricellulin is a tight-junction protein necessary for hearing. *Am. J. Hum. Genet.*, **79**, 1040–1051.
23. Nayak, G., Lee, S.I., Yousaf, R., Edelmann, S.E., Trincot, C., Van Itallie, C.M., Sinha, G.P., Rafeeq, M., Jones, S.M., Belyantseva, I.A. *et al.* (2013) Tricellulin deficiency affects tight junction architecture and cochlear hair cells. *J. clin. invest.*, **123**, 4036–4049.
24. Khan, S.G., Oh, K.S., Shahlavi, T., Ueda, T., Busch, D.B., Inui, H., Emmert, S., Imoto, K., Muniz-Medina, V., Baker, C.C. *et al.* (2006) Reduced XPC DNA repair gene mRNA levels in clinically normal parents of xeroderma pigmentosum patients. *Carcinogenesis*, **27**, 84–94.
25. Khan, S.G., Oh, K.S., Emmert, S., Imoto, K., Tamura, D., Digiovanna, J.J., Shahlavi, T., Armstrong, N., Baker, C.C., Neuburg, M. *et al.* (2009) XPC initiation codon mutation in xeroderma pigmentosum patients with and without neurological symptoms. *DNA repair*, **8**, 114–125.
26. Windhager, L., Bonfert, T., Burger, K., Ruzsics, Z., Krebs, S., Kaufmann, S., Malterer, G., L'Hernault, A., Schilhabel, M., Schreiber, S. *et al.* (2012) Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome Res.*, **22**, 2031–2042.
27. Burger, K., Muhl, B., Kellner, M., Rohrmoser, M., Gruber-Eber, A., Windhager, L., Friedel, C.C., Dolken, L. and Eick, D. (2013) 4-thiouridine inhibits rRNA synthesis and causes a nucleolar stress response. *RNA Biol.*, **10**, 1623–1630.
28. Jonkers, I., Kwak, H. and Lis, J.T. (2014) Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife*, **3**, e02407.
29. Duan, J., Shi, J., Ge, X., Dolken, L., Moy, W., He, D., Shi, S., Sanders, A.R., Ross, J. and Gejman, P.V. (2013) Genome-wide survey of interindividual differences of RNA stability in human lymphoblastoid cell lines. *Sci. Rep.*, **3**, 1318.
30. Schneider-Poetsch, T., Ju, J., Eyler, D.E., Dang, Y., Bhat, S., Merrick, W.C., Green, R., Shen, B. and Liu, J.O. (2010) Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. *Nat. Chem. Biol.*, **6**, 209–217.
31. Rajavel, K.S. and Neufeld, E.F. (2001) Nonsense-mediated decay of human HEXA mRNA. *Mol. Cell. Biol.*, **21**, 5512–5519.
32. Lareau, L.F., Inada, M., Green, R.E., Wengrod, J.C. and Brenner, S.E. (2007) Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, **446**, 926–929.
33. Damgaard, C.K., Kahns, S., Lykke-Andersen, S., Nielsen, A.L., Jensen, T.H. and Kjems, J. (2008) A 5′ splice site enhances the recruitment of basal transcription initiation factors in vivo. *Mol. Cell.*, **29**, 271–278.
34. Kwek, K.Y., Murphy, S., Furger, A., Thomas, B., O'Gorman, W., Kimura, H., Proudfoot, N.J. and Akoulitchev, A. (2002) U1 snRNA associates with TFIIH and regulates transcriptional initiation. *Nat. Struct. Biol.*, **9**, 800–805.
35. Dumesic, P.A., Natarajan, P., Chen, C., Drinnenberg, I.A., Schiller, B.J., Thompson, J., Moresco, J.J., Yates, J.R., 3rd, Bartel, D.P. and Madhani, H.D. (2013) Stalled spliceosomes are a signal for RNAi-mediated genome defense. *Cell*, **152**, 957–968.
36. Bayne, E.H., Portoso, M., Kagansky, A., Kos-Braun, I.C., Urano, T., Ekwall, K., Alves, F., Rappsilber, J. and Allshire, R.C. (2008) Splicing factors facilitate RNAi-directed silencing in fission yeast. *Science*, **322**, 602–606.
37. Wang, Y., Zhu, W. and Levy, D.E. (2006) Nuclear and cytoplasmic mRNA quantification by SYBR green based real-time RT-PCR. *Methods*, **39**, 356–362.
38. Dolken, L. (2013) High resolution gene expression profiling of RNA synthesis, processing, and decay by metabolic labeling of newly transcribed RNA using 4-thiouridine. *Methods Mol. Biol.*, **1064**, 91–100.
39. Schmittgen, T.D. and Livak, K.J. (2008) Analyzing real-time PCR data by the comparative C(T) method. *Nat. Protoc.*, **3**, 1101–1108.
40. Emig, D., Salomonis, N., Baumbach, J., Lengauer, T., Conklin, B.R. and Albrecht, M. (2010) AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. *Nucleic Acids Res.*, **38**, W755–W762.
41. Flicek, P., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.
42. Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B. *et al.* (2013) The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.*, **41**, D64–D69.

Hum Genet (2017) 136:1093–1111
DOI 10.1007/s00439-017-1809-4

REVIEW

Deep intronic mutations and human disease

Rita Vaz-Drago[1] · Noélia Custódio[1] · Maria Carmo-Fonseca[1]

Received: 24 March 2017 / Accepted: 5 May 2017 / Published online: 12 May 2017

Abstract Next-generation sequencing has revolutionized clinical diagnostic testing. Yet, for a substantial proportion of patients, sequence information restricted to exons and exon–intron boundaries fails to identify the genetic cause of the disease. Here we review evidence from mRNA analysis and entire genomic sequencing indicating that pathogenic mutations can occur deep within the introns of over 75 disease-associated genes. Deleterious DNA variants located more than 100 base pairs away from exon–intron junctions most commonly lead to pseudo-exon inclusion due to activation of non-canonical splice sites or changes in splicing regulatory elements. Additionally, deep intronic mutations can disrupt transcription regulatory motifs and non-coding RNA genes. This review aims to highlight the importance of studying variation in deep intronic sequence as a cause of monogenic disorders as well as hereditary cancer syndromes.

Introduction

An average human protein-coding gene is composed of 8–10 short coding pieces termed exons interrupted by approximately 20 times longer non-coding sequences or introns. Introns are a hallmark of eukaryotic evolution and a substantial intron gain has accompanied the origin of metazoa (Irimia and Roy 2014). The intron–exon structure of eukaryotic genes has played a major role in the creation of new genes through exon shuffling (Berk 2016; Long et al. 2003), and the ability to alternatively select different combinations of exons has been critical for generating gene expression diversity in more complex organisms (Keren et al. 2010).

During transcription, the entire sequence information of a gene is copied into a precursor messenger mRNA (pre-mRNA), which includes exons and introns. To form a contiguous coding sequence that can be translated into a protein, introns have to be precisely removed from the pre-mRNA by a process called RNA splicing (Gilbert 1978). Genes producing long non-coding RNAs (lncRNAs) also typically contain an intron–exon structure and are often spliced (Schlackow et al. 2017).

Efficient and precise excision of introns is catalyzed by a highly sophisticated ribonucleoprotein machinery called the spliceosome (Papasaikas and Valcarcel 2016). Intron–exon boundaries are delimited by short consensus sequences at the 5′ (donor) and 3′ (acceptor) splice sites (ss) that mediate recognition by the spliceosome; in addition, spliceosomal components interact with a catalytic adenosine (the branch point) and a polypyrimidine tract (PyT) located between the branch point adenosine and the 3′ss (Fig. 1).

The spliceosome is formed by five small RNAs (the U1, U2, U4, U5, and U6 snRNAs) and more than 200 proteins (Wahl et al. 2009). A subset of these proteins associate with the snRNAs forming functional particles called the U1, U2, U4, U5, and U6 snRNPs. Each snRNP consists of a snRNA molecule associated with a common set of Sm proteins and a variable number of additional specific proteins. Within the spliceosome, the consensus splicing sequences in the pre-mRNA are forced into 3-dimensional arrangements that enable the activation of an RNA catalytic center and trigger the splicing reaction (Hang et al. 2015).

✉ Maria Carmo-Fonseca
carmo.fonseca@medicina.ulisboa.pt

[1] Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Av. Prof. Egas Moniz, 1649-028 Lisbon, Portugal

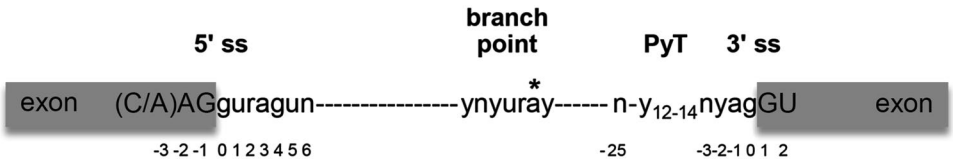


Fig. 1 Human consensus splice site sequences. The most conserved nucleotide sequence is indicated for canonical 5′ (donor) and 3′ (acceptor) splice sites (ss), the branch point (a*) and polypyrimidine tract (PyT). The donor site (5′ss) is defined by the three terminal nucleotides of each exon and the first seven bases of the downstream intron. The acceptor site (3′ss) consists of the first two bases of the exon and the last 26 bases of the upstream intron, including the PyT. The coordinate ranges for the donor and acceptor site positions are [−3, +6] and [−25, +2], with a zero coordinate at the first intronic base of each splice site (Caminsky et al. 2014)

Spliceosome assembly starts with the recognition of the 5′ss by the U1 snRNP (Seraphin and Rosbash 1989). Subsequently, the U2 snRNP associates with the branch point region. However, this interaction requires prior binding of splicing factor 1 (SF1) to the branch point sequence (Berglund et al. 1997; Liu et al. 2001) and recruitment of U2 auxiliary factor U2AF to the 3′ss (Zamore et al. 1992). U2AF is an heterodimer composed of a 65 kDa subunit that contacts the polypyrimidine tract and a 35 kDa subunit that recognizes the AG at the intron 3′ end (Merendino et al. 1999; Wu et al. 1999; Zorio and Blumenthal 1999). The U4/U6.U5 tri-snRNP particle is then recruited, forming the pre-catalytic spliceosome (Will and Luhrmann 2011). After the release of U1 and U4, the adenosine residue at the branch point undergoes a nucleophilic attack on the 5′ss, resulting in the formation of a 2′,5′-phosphodiester bond. This is followed by the second trans-esterification reaction, which consists in the 5′ss-mediated attack on the 3′ss, giving rise to the spliced product and releasing the intron lariat (Will and Luhrmann 2011).

The vast majority of introns are processed by the U1, U2, U4, U5, and U6 snRNP complex, also known as the major spliceosome. Yet, a minority of introns are spliced by a distinct type of snRNP complex called the minor spliceosome (Patel and Steitz 2003). Overall, the major and minor spliceosomes share many common features and the mechanism of splicing is nearly identical. However, the minor spliceosome is composed of four distinct snRNAs termed as U11, U12, U4atac, and U6atac that have a counterpart in U1, U2, U4 and U6, respectively (Hall and Padgett 1996; Tarn and Steitz 1996a, b; Will et al. 1999).

Most of the interactions between pre-mRNA and spliceosomal snRNAs are weak and prone to be modulated by multiple mechanisms that involve the binding of splicing regulatory proteins to the pre-mRNA, the formation of secondary structures in the pre-mRNA, the rate of Pol II transcriptional elongation, and epigenetic modification of the template chromatin [for a recent review see (Naftelberg et al. 2015)]. Depending on the combinatorial effect of factors that either enhance or repress the recognition of consensus sequences by the spliceosome, different splice sites will be selected in the pre-mRNA (Black 2003). The recent combination of large-scale characterization of alternative splicing and genome-wide identification of in vivo binding sites of splicing regulators unraveled the global principles guiding splicing regulation by specific RNA-binding proteins (Barash et al. 2010; Witten and Ule 2011). Typically, splicing regulatory elements are classified as exonic or intronic splicing enhancers or silencers depending on their location and ability to stimulate or inhibit splicing (Wang and Burge 2008).

Alterations in pre-mRNA splicing are increasingly recognized as responsible for monogenic disorders (Chabot and Shkreta 2016; Krawczak et al. 2007; Padgett 2012; Pedrotti and Cooper 2014; Scotti and Swanson 2016; Singh and Cooper 2012; Sterne-Weiler and Sanford 2014), as well as for complex human traits (Heinzen et al. 2008; Yu et al. 2008). Shortly after the discovery of splicing it was found that patients with β-thalassemia failed to produce β-globin (HBB) protein due to a point mutation in an intron that disrupted the normal processing of *HBB* pre-mRNA (Busslinger et al. 1981; Spritz et al. 1981). Current estimates indicate that between 15 and 50% of all monogenic disease-causing mutations affect pre-mRNA splicing (Wang and Cooper 2007) with a large fraction corresponding to point mutations in splice junctions (Krawczak et al. 2007; Stenson et al. 2012). In addition, mutations in the binding sites for splicing regulatory proteins or mutations in the genes encoding these proteins are known to contribute to aberrant splicing and disease phenotypes (Caminsky et al. 2014), with up to 25% of known missense and nonsense disease-causing mutations predicted to alter exonic splicing enhancers or silencers (Cartegni et al. 2002; Sterne-Weiler et al. 2011). Likewise, mutations affecting intronic splicing enhancers or silencers have the potential to result in mis-regulated splicing.

The recent introduction of whole-genome sequencing approaches in clinically oriented screening studies has resulted in the identification of an increasing number of pathogenic variants located deep within introns (i.e., more than 100 base pairs away from exon–intron boundaries). Additionally, Genome Wide Association Studies

have identified many single nucleotide variants located deep within introns with significant association to diseases (Hsiao et al. 2016; Xiong et al. 2015). These findings are fostering a new era of research focused on understanding how variation in deep intronic sequence affects pre-mRNA splicing and contributes to disease phenotypes.

Uncovering deep intronic secrets

For many years the physiological importance of a genomic sequence was primarily associated with its capacity to code for proteins and, therefore, intronic sequences were initially assumed to be largely non-functional. However, several lines of recent evidence argue for intron functionality, as discussed in detail in the following sections.

Intron conservation

Because introns are removed from nascent transcripts during pre-mRNA processing, intronic sequences in genes have been considered as "junk DNA". However, there is a remarkable conservation of many intron positions along with highly conserved sequence elements, implying that at least some intronic features are subject to evolutionary constraints (Hare and Palumbi 2003; Mattick and Gagen 2001; Rogozin et al. 2003). Conserved elements in introns include the consensus splice site sequences, the binding sites for regulatory proteins, the sequences of non-coding RNA genes, as well as additional regions (Kelly et al. 2015).

Several studies revealed that first introns (i.e., introns at the 5′ end of genes) are typically the longest and most conserved (Park et al. 2014). Conservation of the first intron is probably related to the presence of regulatory elements (Gaffney and Keightley 2004) and a specific pattern of chromatin organization (Bieberstein et al. 2012). Additionally, the position of introns that contain RNA genes was also found to be highly conserved (Chorev and Carmel 2013).

Introns as enhancers of transcription

Introns were first shown to increase transcriptional efficiency in transgenic mice (Brinster et al. 1988). Subsequent studies showed that intron-containing genes presented higher levels of transcription when compared to intronless genes in yeast (Juneau et al. 2006), Drosophila (McKenzie and Brennan 1996) and mammalian cells (Brinster et al. 1988; Shabalina et al. 2010). Transcription of mammalian genes relies on a complex communication between promoters and enhancers that are often located a large distance apart in the genome, and recent studies suggest that some promoters work in combination with regulatory sequences located within introns (Stadhouders et al. 2012). For example, expression of the type II collagen α1 (*Col2a1)* gene is dependent on SOX9, a master transcription factor that binds to regulatory regions located in *Col2a1* introns 1 and 6 (Yasuda et al. 2017). Likewise, expression of the vascular endothelial growth factor receptor *Flk1* gene requires a regulatory region located in intron 10 (Becker et al. 2016). Another example, is an enhancer for the Sonic Hedgehog *SHH* gene, which is located 1 Mbp upstream, within an intron of the unrelated *LMBR1* gene (Sagai et al. 2005). The promoter-proximal 5′ splice site was further shown to stimulate transcription independently from splicing, presumably through the binding of U1 snRNP and its interaction with transcription initiation factors (Damgaard et al. 2008; Kwek et al. 2002).

Intron-encoded RNA genes

After splicing, introns initially excised in lariat form are first debranched (Ruskin and Green 1985) and then in most cases rapidly degraded (Sharp et al. 1987). Yet, not all introns are fully degraded, but rather give rise to functional non-coding RNA by-products (Hube and Francastel 2015; Mattick 2001). These include most small nucleolar RNAs (snoRNAs), which are produced from processed introns derived from genes encoding various ribosomal proteins, ribosome-associated proteins, nucleolar and other proteins (Maxwell and Fournier 1995). Remarkably, some genes have no protein-coding capacity and their primary function may be to generate snoRNAs from their introns (Bortolin and Kiss 1998; Tycowski et al. 1996). In addition to snoRNAs, a class of unconventional micro RNAs (miRNAs) is also produced from introns. In this case, pre-miRNA-like hairpins are generated by the spliceosome followed by lariat-debranching and exosome mediated trimming (Flynt et al. 2010). These atypical miRNA precursors are called mirtrons due to their location in introns from protein coding and non-coding genes (Berezikov et al. 2007; Okamura et al. 2007; Valen et al. 2011).

Alternative splicing and intron retention

Alternative splicing increases transcriptome and proteome diversity by generating multiple mRNA isoforms from a single gene. A pre-mRNA molecule can be alternatively spliced through exon skipping, alternative splice site selection, and intron retention (Black 2003; Chen and Manley 2009). For many years, intron retention in mature mRNAs was considered a consequence of mis-splicing, as intron-containing mRNAs are often targeted for degradation by the exosome in the nucleus or nonsense-mediated decay in the cytoplasm (Gudipati et al. 2012; Jaillon et al. 2008;

Roy and Irimia 2008). However, recent transcriptomic analysis revealed that many introns are actively retained in polyadenylated transcripts and contribute to downregulate gene expression (Wong et al. 2013; Yap et al. 2012). Yet, transcripts with retained introns are not necessarily short-lived. For example, in the mouse brain, mRNAs containing certain introns are stably accumulated in the nucleus, but in response to a stimulus, these molecules are spliced and acutely transported to the cytoplasm (Mauger et al. 2016; Naro et al. 2017). Similarly, nuclear accumulation of stable transcripts containing retained introns was detected at specific stages during spermatogenesis (Naro et al. 2017). The term "detained" introns has been proposed to describe this class of introns that are transiently retained in nuclear transcripts, but can still be spliced (Boutz et al. 2015). Retained introns can additionally affect gene expression by other mechanisms. Namely, the first intron of the *ZEB2* pre-mRNA contains an internal ribosome entry site, so that retention of this intron allows more efficient translation (Beltran et al. 2008), and retention of the third intron in the rat Ceacam6-L pre-mRNA generates a novel protein isoform in male germ cells (Kurio et al. 2008).

Non-canonical splicing

Recent developments in transcriptome sequencing and analysis have revealed a remarkable prevalence of unconventional or non-canonical splicing mechanisms ranging from recognition of atypical splice sites to changes in the usual order of splicing [for a recent review see (Sibley et al. 2016)].

Many sequences similar to the consensus motifs of canonical splice sites are present throughout introns. These sequences are known as cryptic, non-canonical or pseudo splice sites. What determines preference for a bona fide versus a pseudo splice site is still unclear, particularly after the finding that actual splice site sequences can be extremely diverse (Roca et al. 2003, 2013). Indeed, over 9000 sequence variants have been found in the −3 to +6 region of human 5′ splice sites (Roca et al. 2012), challenging the dogma that spliceosome recognition relies primarily on consensus sequences at exon–intron boundaries. Moreover, cryptic splice sites are often used when a natural splice site is mutated, further arguing that splice site recognition is not intrinsically defined by any given sequence (Roca et al. 2003, 2013). Most likely, bona fide splice site selection results from the combinatorial effect of proteins such as SR and hnRNP proteins that bind to the pre-mRNA and either stabilize spliceosome interactions or inhibit the recruitment of spliceosomal components (Dreyfuss et al. 2002; Liu et al. 1998).

Intronic sequences that are flanked by non-canonical splice sites and are normally not observed in spliced mRNAs are referred to as pseudo-exons or cryptic exons. Compared to genuine exons, pseudo-exons tend to have less splicing enhancer and more splicing silencer motifs (Corvelo and Eyras 2008; Sironi et al. 2004; Wang et al. 2004; Zhang and Chasin 2004). Pseudo-exons often derive from transposable elements, in particular from antisense *Alu* sequences (Keren et al. 2010).

A non-canonical mechanism of intron removal referred to as recursive splicing was first detected in long Drosophila pre-mRNAs (Burnette et al. 2005; Hatton et al. 1998). Recently, recursive splicing was also observed in long introns of mammalian brain-specific transcripts (Sibley et al. 2015). These introns contain a cryptic site termed a recursive splice site or a 'zero-length exon' consisting in a combination of 3′ and 5′ splice sites that allow an intron to be spliced in multiple consecutive steps (Duff et al. 2015; Sibley et al. 2015). In this process, the 3′ splice site is used to splice the upstream part of the intron, which reconstitutes a 5′ splice site that is then used to splice the downstream part. Evidence for multi-step recursive splicing of dystrophin pre-mRNAs in human skeletal muscle cells has also been reported (Gazzoli et al. 2016).

In some cases, the pre-mRNA splicing reaction does not follow its canonical order, but rather occurs in a reversed orientation that links a downstream 5′ (donor) site to an upstream 3′ (acceptor) site to produce a circular RNA [for recent reviews see (Chen 2016; Salzman 2016)]. To date, thousands of circular non-coding RNAs generated by "backsplicing" of transcripts from protein-coding genes have been reported. Circularization, which results, for example, from covalently joining the two ends of a single exon, can be favored by the presence of inverted repeats, such as *Alu* elements, in the flanking introns (Dong et al. 2016; Jeck et al. 2013; Liang and Wilusz 2014; Wilusz 2015). Intronic circular RNAs have further been detected resulting from intron lariats that are resistant to de-branching due to C-rich motifs surrounding the branch point (Zhang et al. 2013). Although many circular RNA species appear to result from splicing errors, some may function as modulators of gene expression (Chen 2016; Salzman 2016).

Deep intronic mutations as cause of human disease

To date, mutations in deep intronic regions have been documented in multiple diseases. In the following sections we review reported mutations and discuss the mechanism by which they alter gene expression. We reviewed 117 studies published between 1983 and 2016 describing 185 intronic mutations located at least 100 bp from the nearest canonical splice site, across 77 different disease genes.

Inclusion of pseudo-exons

Pseudo-exon inclusion is now considered a more frequent cause of disease than previously thought (Dhir and Buratti 2010; Romano et al. 2013). This aberrant process can be triggered by intronic mutations that activate non-canonical splice sites. The more common mechanism involves a mutation that creates a novel donor splice site and activates a pre-existing non-canonical acceptor splice site (Fig. 2a). Less frequently the mutation creates a novel acceptor splice site (Fig. 2b). Alternatively, inclusion of a pseudo-exon results from mutations that either create or disrupt splicing enhancer or silencer elements, respectively (Fig. 3a, b). Disease-associated pseudo- or cryptic exons range in size from 30 to 344 base pairs (Fig. 4) and about half of them are derived from transposable elements, particularly *Alu* elements (Vorechovsky 2010). The appearance of a pseudo-exon generally disrupts the reading frame introducing a premature termination codon that targets the mutant mRNA for degradation by nonsense-mediated decay (NMD) (Popp and Maquat 2013).

Pseudo-exon inclusion was first reported in β-Thalassemia patients (Dobkin et al. 1983; Treisman et al. 1983). A T>G mutation located 705 bp downstream of the *HBB* middle exon created a new donor splice site and activated an acceptor splice site present within the second intron (Dobkin et al. 1983). Several additional deep intronic mutations leading to pseudo-exon inclusion have since been identified in patients affected by multiple disorders (Tables 1, 2, 3).

The longer a gene the more likely it is to be affected by pathogenic mutations (Lopez-Bigas et al. 2005). It is, therefore, not surprising that numerous deep intronic mutations have been described in particularly long genes such as those associated with neurofibromatosis (Cunha et al. 2016) and Duchenne muscular dystrophy (Beroud et al. 2004; Gonorazky et al. 2016; Gurvich et al. 2008; Trabelsi et al. 2014). Remarkably, deep intronic mutations that promote inclusion of a pseudo-exon have been described in several hereditary tumor syndromes (Tables 1, 2, 3). These include neurofibromatosis types 1 and 2 (Castellanos et al. 2013; Svaasand et al. 2015), melanoma (Harland et al. 2001), ataxia-telangiectasia (Coutinho et al. 2005), retinoblastoma (Dehainault et al. 2007), Lynch syndrome (Clendenning et al. 2011), breast cancer (Anczukow et al. 2012), and familial adenomatous polyposis (Spier et al. 2012). In the majority of these cases the mutant mRNA species are degraded by NMD.

Although point mutations are most frequent, small intronic deletions have also been reported to trigger pseudo-exon inclusion. For example, a 18 bp deletion within the intron 37 of the *DMD* gene was found associated with insertion of a 77 nt pseudo-exon in the mRNA (Bovolenta et al. 2008).

A splicing enhancer created de novo within an intronic region may be sufficient to promote recognition by the spliceosome leading to pseudo-exon inclusion (Fig. 3a). For example, an intronic mutation was identified in a Becker muscular dystrophy patient that created two splicing enhancer sites in intron 26 of the dystrophin gene resulting

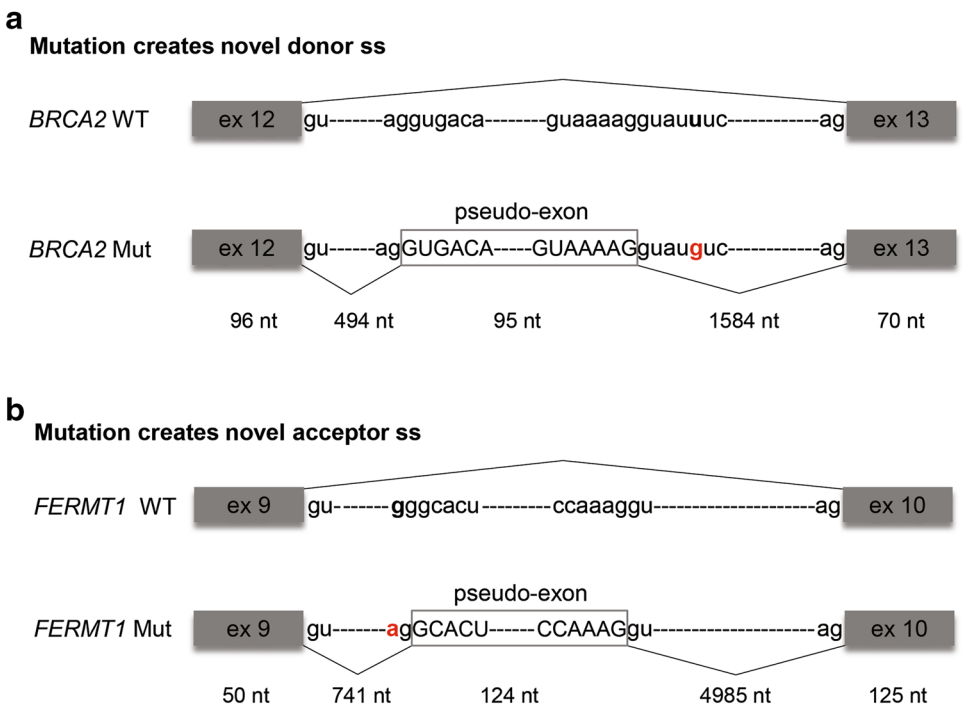


Fig. 2 Pseudo-exon inclusion triggered by mutations that activate non-canonical splice sites. **a** Representative example of a mutation that creates a novel donor splice site (ss) and activates an upstream acceptor splice site; this leads to inclusion of a 95 nucleotide (nt) intronic sequence (pseudo-exon) in the *BRCA2* mRNA (Anczukow et al. 2012); **b** representative example of a mutation that creates a novel acceptor splice site and activates a downstream donor site; this leads to inclusion of a 124 nt intronic sequence (pseudo-exon) in the *FERMT1* mRNA (Chmel et al. 2015)

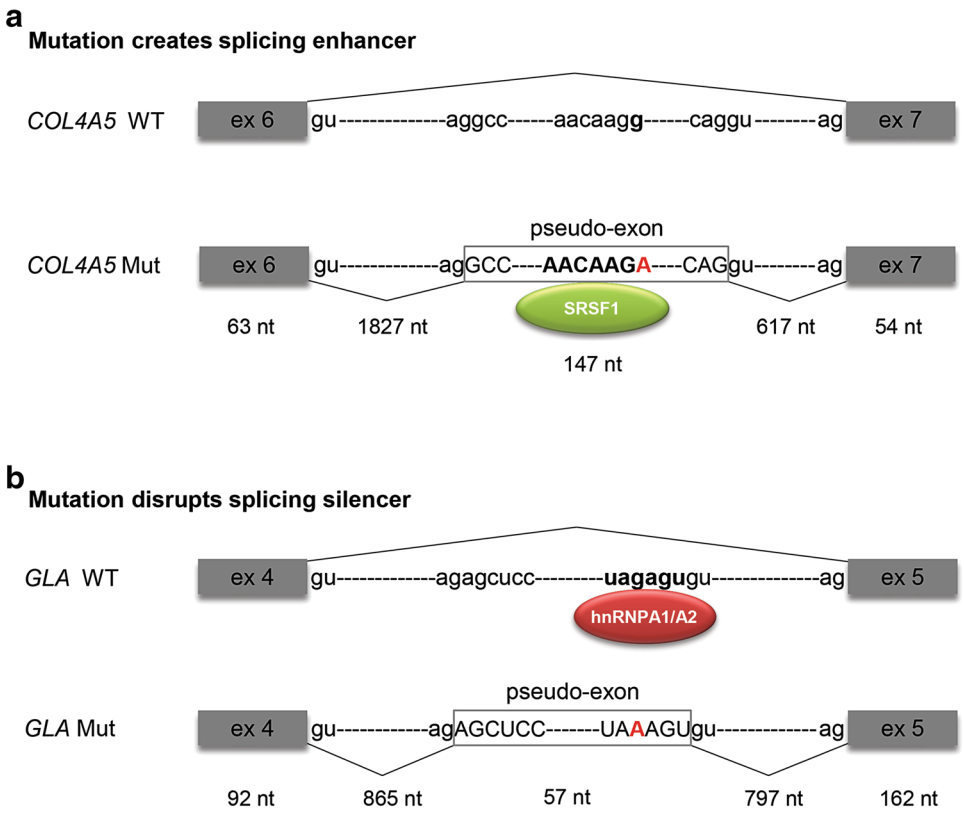


Fig. 3 Pseudo-exon inclusion triggered by mutations that alter splicing regulatory elements. **a** Representative example of a mutation that creates a novel binding site for SRSF1, thus activating a splicing enhancer element; this leads to inclusion of a 147 nt pseudo-exon in the *COL4A5* mRNA (King et al. 2002); **b** representative example of a mutation that disrupts a binding site for hnRNPA1/A2, thus inactivating a splicing silencer element; this leads to inclusion of a 57 nt pseudo-exon in the *GLA* mRNA (Palhais et al. 2016)

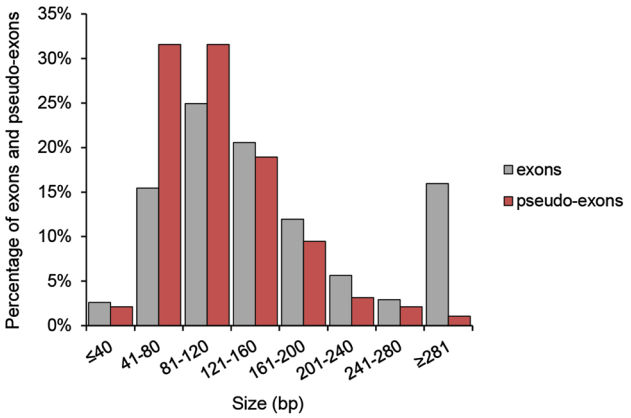


Fig. 4 Size distribution of pseudo-exons. Pseudo-exons referred to in Tables 1, 2 and 3 were size-distributed in 40 bp intervals. For comparison, the size distribution of authentic exons is indicated. Size and frequency of authentic human exons (*h19*) were calculated using BED files downloaded from the UCSC *Table Browser*

in the insertion of a new cryptic exon in the mRNA (Trabelsi et al. 2014). A deep intronic mutation was also identified in a patient with X-linked form of Alport syndrome that created a new enhancer sequence triggering the inclusion of a cryptic exon (King et al. 2002).

Alternatively, inclusion of cryptic exons can be induced by mutations that inactivate repressive sequences or secondary structures (Buratti et al. 2007; Greer et al. 2015; Highsmith et al. 1994) (Fig. 3b). For example, a prevalent deep intronic mutation (c.639+919G>A) in the lysosomal α-galactosidase A (*GLA*) gene responsible for Fabry disease disrupts an hnRNP A1 and hnRNP A2/B1-binding splicing silencer motif. This allows binding of U1 snRNP to an overlapping cryptic 5′ss resulting in pseudo-exon inclusion (Palhais et al. 2016). Moreover, disruption of intronic U1 snRNP binding sites was found to trigger pseudo-exon inclusion in patients with ataxia-telangiectasia (Pagani et al. 2002) and Laron syndrome (Akker et al. 2007). Possibly, splicing-independent binding of U1 snRNP to intronic regions suppresses both pseudo-exon inclusion and premature cleavage and polyadenylation from cryptic polyadenylation signals located in introns (Kaida et al. 2010). In some cases, the deep intronic mutation leads to the appearance of more than one aberrantly spliced mRNA isoforms. For example, two 46, XY sisters with complete androgen insensitivity syndrome had a deep intronic mutation upstream of exon 7 in the androgen receptor (*AR*) gene. The mutation (c.2450-118A>G) created a de novo 5′ss and activated a novel 3′ss 84 bp upstream, leading to pseudo-exon inclusion in approximately half of the mRNAs. Another half of the cryptically spliced mRNAs showed inclusion of the entire 202 nt intronic fragment downstream of the novel 3′ss, most likely through creation of a novel exonic splicing enhancer motif specific for the binding of

Table 1 Human diseases caused by deep intronic mutations that create a new donor splice site leading to inclusion of a pseudo-exon

Disease (OMIM)	Gene	Deep intronic variant	References	Additional reports
Monogenic diseases				
β-Thalassemia7 (6713985)	*HBB*	IVS2+705T>G	Dobkin et al. (1983)	Treisman et al. (1983), Cheng et al. (1984)
Gyrate atrophy of choroid and retina with or without ornithinemia (258870)	*OAT*	IVS3+303C>G	Mitchell et al. (1991)	
Cystic Fibrosis (2197700)	*CFTR*	3849+10KbC>T	Highsmith et al. (1994)	Abeliovich et al. (1992), Chillon et al. (1995), Friedman et al. (1999), Monnier et al. (2001), Costantino et al. (2013)
Mucopolysaccharidosis II (309900)	*IDS*	1131-133A>G	Rathmann et al. (1996)	
Hyperphenylalaninemia, BH4-deficient, C (261630)	*QDPR*	IVS3+2552A>G	Ikeda et al. (1997)	
Maple syrup urine disease, type II (248600)	*DBT*	IVS8-550A>G	Tsuruta et al. (1998)	
Mucopolysaccharidosis type VII (253220)	*GUSB*	IVS8+0.6kbdelTC	Vervoort et al. (1998)	
Tuberous sclerosis (613254)	*TSC2*	IVS8+281C>T	Mayer et al. (2000)	
Congenital cataracts facial dysmorphism neuropathy syndrome (604168)	*CTDP1*	IVS6+389C>T	Varon et al. (2003)	
Central core disease (117000)	*RYR1*	IVS100+2990A>G	Monnier et al. (2003)	
Hyper IgM Syndrome, type 1 (308230)	*CD40L*	IVS3-915A>T	Lee et al. (2005)	
Chronic granulomatous disease (306400)	*CYBB*	IVS6-1157A>G	Rump et al. (2006)	Noack et al. (2001)
Leber congenital amaurosis 10 (611755)	*CEP290*	c.2991+1655A>G	den Hollander et al. (2006)	Ruan et al. (2017)
Schwartz-Jampel Syndrome (255800)	*HSPG2*	c.574+481C>T	Stum et al. (2006)	
Congenital disorder of glycosylation, type Ia (212065)	*PMM2*	c.639-15479C>T	Schollen et al. (2007)	Vega et al. (2009)
Afibrinogenemia (202400)	*FGG*	IVS6-320A>T	Spena et al. (2007)	Plate et al. (2009)
Methylmalonic aciduria, mut(0) type (251000)	*MUT*	IVS11+3691C>A	Rincon et al. (2007)	
Propionicacidemia (606054)	*PCCB*	IVS6+462A>G	Rincon et al. (2007)	
Leigh syndrome (256000)	*NDUFS7*	c.17-1167C>G	Lebon et al. (2007)	
Mitochondrial trifunctional protein deficiency (609105)	*HADHB*	IVS7+614A>G	Purevsuren et al. (2008)	
Deafness, autosomal dominant 22 (606346)	*MYO6*	IVS23+2321T>G	Hilgert et al. (2008)	
Polycystic kidney and hepatic disease (263200)	*PKHD1*	IVS46+1653A>G	Michel-Calemard et al. (2009)	
Amyotrophic lateral sclerosis 1 (105400)	*SOD1*	c.358–304C>G	Valdmanis et al. (2009)	
Niemann-Pick disease, type C1 (250227)	*NPC1*	c.1554-1009G>A	Rodriguez-Pascau et al. (2009)	
Retinitis pigmentosa 11 (600138)	*PRPF31*	c.1374+654C>G	Rio Frio et al. (2009)	
Duchenne muscular dystrophy (310200)	*DMD*	c.3787-843C>A c.9807+2714C>T	Takeshima et al. (2010)	Beroud et al. (2004), Deburgrave et al. (2007), Bovolenta et al. (2008), Ikezawa et al. (1999), Gurvich et al. (2008)
Becker muscular dystrophy (30376)	*DMD*	c.9225-285A>G	Takeshima et al. (2010)	Tuffery-Giraud et al. (2003), Gurvich et al. (2008)
5-fluorouracil toxicity (274270)	*DPYD*	c.1129–5923C>G	van Kuilenburg et al. (2010)	

Table 1 continued

Disease (OMIM)	Gene	Deep intronic variant	References	Additional reports
Gitelman's Syndrome (263800)	*SLC12A3*	c.1670-191C>T c.2548+253C>T	Lo et al. (2011)	Nozu et al. (2009)
Megalencephalic leukoencephalopathy with subcortical cysts (604004)	*MLC1*	c.895-226T>G	Mancini et al. (2012)	
Retinitis pigmentosa (300424)	*OFD1*	IVS9+706A>G	Webb et al. (2012)	
Coffin–Lowry syndrome (303600)	*RPS6KA3*	c.1228-279T>G	Schneider et al. (2013)	
Hyperinsulinemic hypoglycemia, familial, 1 (256450)	*ABCC8*	c.1333-1013A>G	Flanagan et al. (2013)	
Inherited growth-hormone insensitivity (604271)	*GHR*	c.618+792A>G	Walenkamp et al. (2013)	David et al. (2007), Metherell et al. (2001)
Hemophilia A (306700)	*F8*	c.5998+530C>T	Pezeshkpoor et al. (2013)	Bagnall et al. (1999), Castaman et al. (2011), Inaba et al. (2013)
Stargardt disease 1 (248200)	*ABCA4*	c.5196+1056A>G	Braun et al. (2013), Schulz et al. (2017)	Braun et al. (2013), Bauwens et al. (2015), Bax et al. (2015)
Marfan syndrome (154700)	*FBN1*	c.6872-961A>G	Gillis et al. (2014)	Guo et al. (2008)
Gorlin syndrome (109400)	*PTCH1*	c.2561-2057A>G	Bholah et al. (2014)	
Menkes disease (309400)	*ATP7A*	c.2406+1117A>G	Yasmeen et al. (2014)	
Miyoshi muscular dystrophy 1(254130)	*DYSF*	c.4886+1249G>T	Dominov et al. (2014)	
Hyperinsulinemic hypoglycemia, familial, 4 (609975)	*HADH*	c.636+471G>T	Camtosun et al. (2015)	Flanagan et al. (2013)
Ocular albinism (300500)	*GPR143*	c.659-131T>G	Naruto et al. (2015)	Vetrini et al. (2006)
Ushar syndrome type II (276901)	*USH2A*	c.9959-4159A>G c.5573-834A>G c.8845+628C>T	Liquori et al. (2016)	
Complete androgen insensitivity syndrome (300068)	*AR*	c.2450-118A>G	Kansakoski et al. (2016)	
Duchenne muscular dystrophy (310200)	*DMD*	deletion 18 bp[a]	Bovolenta et al. (2008)	
Hereditary tumor syndromes				
Retinoblastoma (180200)	*RB1*	IVS23-1398A>G	Dehainault et al. (2007)	
Lynch syndrome (120435)	*MSH2*	c.212-478T>G	Clendenning et al. (2011)	
Breast cancer (114480)	*BRCA2*	c.6937+594T>G	Anczukow et al. (2012)	Balz et al. (2002)
Familial Adenomatous Polyposis (175100)	*APC*	c.532-941G>A c.1408+731C>T c.1408+735A>T	Spier et al. (2012)	
Neurofibromatosis type 2 (101000)	*NF2*	c.1447-240T>A	Castellanos et al. (2013)	De Klein et al. (1998)
Neurofibromatosis type 1 (162200)	*NF1*	c.288+1137C>T	Svaasand et al. (2015)	
Ataxia-telangiectasia (208900)	*ATM*	IVS28-159A>G[a]	Coutinho et al. (2005)	McConville et al. (1996)

[a] The mutation leads to activation of both donor and acceptor non-canonical splice sites

Table 2 Human diseases caused by deep intronic mutations that create a new acceptor splice site leading to inclusion of a pseudo-exon

Disease (OMIM)	Gene	Deep intronic variant	Reference	Additional reports
Monogenic diseases				
Alport Syndrome (301050)	*COL4A3*		Knebelmann et al. (1995)	
	COL4A5	IVS29+2733A>G	King et al. (2002)	
Choroideremia (303100)	*CHM*	314+10127T>A	van den Hurk et al. (2003)	
Ornithine transcarbamylase deficiency (311250)	*OTC*	c.540+265G>A	Ogino et al. (2007)	
Myopathy with lactic acidosis, hereditary (255125)	*ISCU*	IVS5+382G>C	Kollberg et al. (2009)	Olsson et al. (2008), Mochel et al. (2008)
Duchenne muscular dystrophy (310200)	*DMD*	c.5326-215T>G	Gonorazky et al. (2016)	Yagi et al. (2003), Gurvich et al. (2008), Beroud et al. (2004), Deburgrave et al. (2007), Bovolenta et al. (2008)
Becker muscular dystrophy (30376)	*DMD*	c.93+5590T>A	Takeshima et al. (2010)	Tuffery-Giraud et al. (2003)
Werner syndrome (277700)	*WRN*	c.3234-160A>G	Friedrich et al. (2010)	Masala et al. (2007)
Lesch–Nyhan syndrome (300322)	*HPRT*	g.36221T>A c.5998+941G>A	Corrigan et al. (2011)	
Limb-girdle muscular dystrophy type 2A (253600)	*CAPN3*	c.1782+1072G>C	Blazquez et al. (2013)	
Optic atrophy plus syndrome (125250)	*OPA1*	c.610+360G>A c.610+364G>A	Bonifert et al. (2016) Bonifert et al. (2014)	
Kindler syndrome (173650)	*FERMT1*	c.1139+740G>A	Chmel et al. (2015)	
Pompe disease (232300)	*GAA*	c.2190-345A>G	Bergsma et al. (2016)	

Table 3 Human diseases caused by deep intronic mutations that interfere with splicing regulatory elements leading to inclusion of a pseudo-exon

Disease (OMIM)	Gene	Deep intronic variant	References	Additional reports
ESE creation				
Monogenic diseases				
Alport Syndrome (301050)	*COL4A5*	IVS6+1873G>A	King et al. (2002)	
Propionicacidemia (606054)	*PCCA*	IVS14-1416A>G	Rincon et al. (2007)	
Afibrinogenemia (202400)	*FGB*	c.115-600A>G	Davis et al. (2009)	
Homocystinuria, type cblE (236270)	*MTRR*	c.903+469T>C	Homolova et al. (2010)	
Becker muscular dystrophy (300376)	*DMD*	c.3603+2053G>C	Trabelsi et al. (2014)	
Complete androgen insensitivity syndrome (300068)	*AR*	c.2450-118A>G	Kansakoski et al. (2016)	
ESS disruption				
Monogenic diseases				
Cystic Fibrosis (219700)	*CFTR*	c.1002–1110_1113delTAAG	Faa et al. (2009)	Nathan et al. (2012), Costa et al. (2011), Straniero et al. (2016)
Fabry disease (301500)	*GLA*	c.639+919G>A	Palhais et al. (2016)	Ishii et al. (2002), Ferri et al. (2016)
Hereditary tumor syndrome				
Ataxia-telangiectasia (208900)	*ATM*	IVS20-579_IVS20-576delGTAA	Pagani et al. (2002)	

SRSF1, a member of the SR protein family. As this motif and 5′ss overlap, SRSF1 and U1 snRNP probably compete with each other for binding to the pre-mRNA, thus resulting in either skipping or inclusion of the cryptic 5′ss (Kansakoski et al. 2016).

Competition with natural splice sites

Most deep intronic mutations have no effect on canonical splice sites. Yet, some mutations that create a new splice site interfere with recognition of natural splice sites (Fig. 5), as

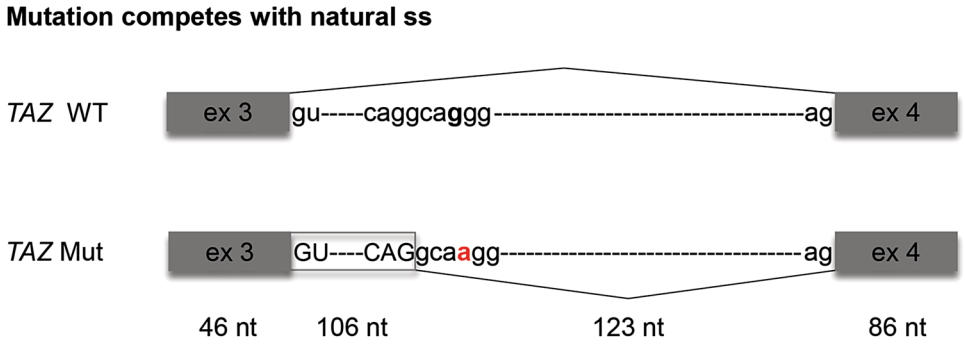


Fig. 5 Competition with natural splice sites. Representative example of a mutation that creates a new donor splice site (ss); this leads to inclusion of an extra 106 nt sequence at the end of exon 3 of the *TAZ* gene (Sakamoto et al. 2001)

reported in the *CDKN2A*, *TAZ*, *PRPF31*, *COL2A1, ATP7A*, *GBE1*, *TCIRG1* and *GAA* genes (Table 4). For example, a mutation deep in intron 2 of *CDKN2A* gene (IVS2-105A>G) created a false GT splice donor site 105 bases 5′ of exon 3 resulting in aberrant splicing of the mRNA (Harland et al. 2001). Weakening of canonical splice sites is frequently observed when deep intronic mutations are less than 150 bp away from the natural exon–intron junctions.

Disruption of transcription regulatory motifs

Multiple cases of genetic diseases caused by deep intronic mutations that disrupt transcription regulatory motifs have been identified. For example, the first intron of the *MPZ* gene contains binding sites for transcription factors SOX10 and EGR2, which are implicated in the regulation of *MPZ* expression (Antonellis et al. 2010). The MPZ protein is required for proper myelination and several coding and splice site mutations in the *MPZ* gene have been described as cause of Charcot-Marie-Tooth disease type 1B, a demyelinating peripheral neuropathy. A patient affected by this disease had a deep intronic rare variant (c.126-1086T>A) that altered a highly conserved nucleotide in the SOX10-binding site and decreased its enhancer activity (Antonellis et al. 2010).

A transcription enhancer element that binds the transcription regulators CTCF and CEBPB has been identified within the first intron of the *FOXF1* gene. A 800 bp deletion affecting this region was shown to abolish the binding of CTCF and CEBPB, and to inhibit *FOXF1* expression (Szafranski et al. 2013).

A deletion spanning up to 2094 bp within intron 1A of the *COL6A2* gene was identified in a Bethlem myopathy patient (Bovolenta et al. 2010). This intronic deletion was found to occur in compound heterozygosity (*trans*) with a small deletion in exon 28. RNA studies showed monoallelic transcription of the *COL6A2* gene, suggesting transcriptional impairment of the intronic mutated allele (Bovolenta et al. 2010).

Preaxial polydactyly, one of the most frequent congenital hand malformations in humans, is associated with point mutations in a highly conserved 800 bp region located deeply within intron 5 of the *LMBR1* gene (Albuisson et al. 2011; Furniss et al. 2008; Gurnett et al. 2007; Lettice et al. 2003). This region, located approximately 8 kb downstream of exon 5, is called ZRS (ZPA regulatory sequence) and has been shown to be essential for proper limb development (Lettice et al. 2003). Like in humans, ZRS point mutations in mice cause supernumerary preaxial digits. ZRS is a transcription regulatory element that controls expression of the

Table 4 Human diseases caused by deep intronic mutations that compete with natural splice sites leading to cryptic splicing

Disease (OMIM)	Gene	Deep Intronic Variant	References
Monogenic diseases			
Barth syndrome (302060)	*TAZ*	IVS3+110G>A	Sakamoto et al. (2001)
Retinitis pigmentosa 11 (600138)	*PRPF31*	c.1374+654C>G	Rio Frio et al. (2009)
Stickler syndrome, type I (108300)	*COL2A1*	c.1527+104T>G	Richards et al. (2012)
Menkes disease (309400)	*ATP7A*	c.2406+1117A>G	Yasmeen et al. (2014)
Adult polyglucosan body disease (232500)	*GBE1*	IVS15+5289_5297deinsTGTTTTTTACATGACAGGT	Akman et al. (2015)
Autosomal recessive osteopetrosis (259700)	*TCIRG1*	c.1887+146G>A c.1887+142T>A	Palagano et al. (2015)
Pompe disease (232300)	*GAA*	c.2190-345A>G	Bergsma et al. (2016)
Hereditary tumor syndromes			
Melanoma, cutaneous malignant, 2 (155601)	*CDKN2A*	IVS2-105A>G	Harland et al. (2001)

Shh gene located at a distance of 1 Mb from *Lmbr1* intron 5 (Albuisson et al. 2011).

Inactivation of non-coding RNA genes

Genetic variants have been reported to cause disease through inactivation of intron-encoded RNA genes. Point mutations in the *RNU4ATAC* gene were identified in patients affected by the developmental disorder Taybi-Linder Syndrome (TALS) or Microcephalic osteodysplastic primordial dwarfism type 1 (MOPD1) (Edery et al. 2011; He et al. 2011). The *RNU4ATAC* gene, which codes for the minor spliceosomal U4atac snRNA, is located within intron 2 of the protein-coding *CLASP1* gene, 682–556 bp upstream of exon 3 (Edery et al. 2011). Like other snRNAs, the U4atac has a complex three-dimensional structure including 2 stem-loops, two stems, and a Sm-protein binding region. Most TALS/MOPD1 mutations cluster in the 5′ stem–loop and are predicted to disrupt the snRNA secondary structure (He et al. 2011). Consistent with loss-of-function of the mutant snRNA, higher levels of unspliced U12-type introns were detected in patient-derived fibroblasts. An additional mutation (g.124G>A) located in the U4atac Sm protein binding site was further shown to reduce significantly the expression of U4atac, and four other mutations (g.30G>A, g.50G>A, g. 50G>C, g.51G>A) located in the 5′ stem-loop decreased the binding of U4atac snRNP proteins NHP2L1 and PRPF31 (Jafarifar et al. 2014). A more recent study reported additional mutations in the *RNU4ATAC* gene as the cause of a distinct developmental disorder, Roifman syndrome (Merico et al. 2015).

Most individuals with Prader–Willi syndrome, a neurodevelopmental disorder, have abnormal expression of genes located on their paternal chromosome 15. The same type of genetic defect confers Angelman syndrome when inherited from the mother. The imprinted domain on human chromosome 15 harbors the *SNRPN* locus, which produces long transcripts with multiple introns that contain snoRNA genes (Runte et al. 2001), and there is evidence indicating that loss of expression of these C/D box snoRNAs contributes to the disease phenotype (Gallagher et al. 2002; Sahoo et al. 2008). Several additional studies suggest a role for snoRNAs in cancer (Williams and Farzaneh 2012). In particular, the SNORD50A and SNORD50B RNAs, which are encoded by genes located within introns 4 and 5 of the long intergenic non-protein coding *SNHG5* gene, are recurrently loss in many types of cancer (Siprashvili et al. 2016). Additionally, a homozygous *SNORD50A* 2-bp deletion which leads to a decrease in SNORD50A transcript levels has been identified in prostate and breast cancer patients (Dong et al. 2008, 2009). The SNORD50A and SNORD50B RNAs bind directly to K-Ras and their loss triggers hyperactivation of Ras-ERK1/ERK2 signaling thus contributing to oncogenesis (Siprashvili et al. 2016).

Genomic rearrangements

Chromosome rearrangements resulting in gene fusions are often detected in cancer cells (Mertens et al. 2015). In most fusions, recombination occurs within introns and gives rise to the expression of chimeric proteins. However, chimeric mRNAs identical to those transcribed from fusion genes have been detected in low abundance in healthy tissues (Janz et al. 2003). Such chimeric mRNAs could be produced by trans-splicing, a process that joins exons from two distinct precursor transcripts (Konarska et al. 1985; Solnick 1985). Although there is still limited evidence for RNA trans-splicing in mammalian cells, it has been proposed that trans-splicing can occur between pre-mRNAs for *JAZF1* and *JJAZ1* in normal cells (Li et al. 2008). Another putative trans-spliced mRNA is the fusion transcript *SLC45A3-ELK4* in which *SLC45A3* exon 1 is fused to *ELK4* exon 2 leading to the expression of a novel protein that was found expressed in normal and benign cancer prostate cells (Rickman et al. 2009). RNA trans-splicing has also been experimentally manipulated as a tool to correct endogenous mutant pre-mRNAs for human gene therapy (Berger et al. 2016; Puttaraju et al. 1999).

Rearrangements involving deep intronic regions have been described, although rarely, in association with monogenic diseases. A major genomic rearrangement was detected in a boy affected with Duchenne muscular dystrophy that consisted in a 90 kb insertion of non-coding chromosome 4 into intron 43 of the dystrophin gene (Baskin et al. 2011). Analysis of mRNA from a muscle biopsy revealed the presence of a cryptic exon originating from chromosome 4 inserted between *DMD* exons 43 and 44 (Baskin et al. 2011). Three complex genomic rearrangements involving the *DMD* gene also leading to variable inclusion of pseudo-exons in the mRNA were further described in Duchenne muscular dystrophy patients (Khelifi et al. 2011; Madden et al. 2009). Additionally, a sequence rearrangement 2.4 kb downstream from exon 11 of the *DMD* gene was found to be the cause of X-Linked Dilated Cardiomyopathy (Ferlini et al. 1998). This rearrangement involving the intron 11 of the *DMD* gene leads to the activation of a cryptic splice site and inclusion of an *Alu*-like sequence in the mature RNA, triggering this transcript for degradation (Ferlini et al. 1998).

An interchromosomal rearrangement affecting a gene involved in severe intellectual disability (*IQSEC2*) has also been identified. In this case, a duplication event occurred on chromosome 4, which was then inserted into chromosome X. More precisely, the last six exons of the *TENM3*

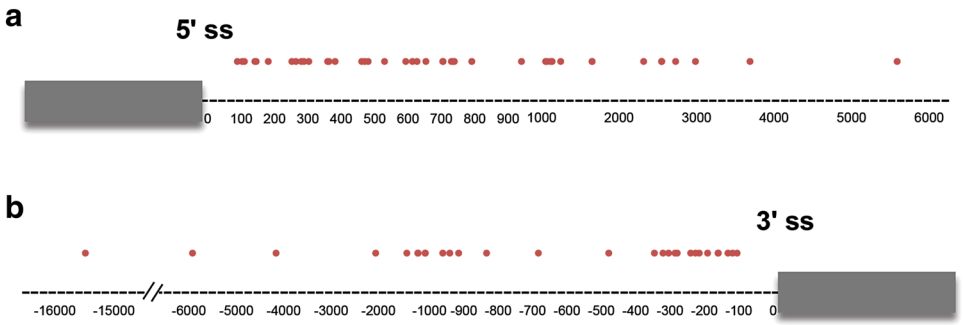


Fig. 6 Distribution of deep intronic mutations across introns. The location of all mutations referred to in Tables 1, 2, 3, 4 is indicated relative to natural 5′ (**a**) and 3′ (**b**) splice sites (5′ss, 3′ss)

gene were inserted in inverted orientation into intron 2 of the *IQSEC2* gene resulting in an in-frame fusion mRNA that escaped NMD (Gilissen et al. 2014).

Conclusion

Despite major advances in clinical genetic analysis introduced by the application of next-generation sequencing strategies, approximately half of the patients remain without a precise genetic diagnosis, which represents a significant limitation for clinical care. In this review we highlight that DNA intronic variants located throughout introns (Fig. 6) can be the cause of human disease and should be investigated when first line approaches such as next-generation sequencing-based gene panels, whole-exome sequencing, microarray and multiplex ligation-dependent probe amplification-based deletion/duplication analysis fail to identify a causative mutation.

To find novel deep intronic mutations and determine their pathogenicity it is crucial to combine sequencing of intronic regions with studies addressing the mRNA molecules produced in affected tissue from patients. This can be done by conventional RT-PCR analysis and sequencing of cDNA products, or by direct RNA-seq analysis. Examination of the patient's transcriptome has the advantage of rapidly detecting the presence of abnormal splicing isoforms (Gonorazky et al. 2016). Reduced levels of mutant transcripts are normally indicative of a disease-causing mutation that either disrupts normal splicing and targets abnormal mRNAs for degradation or inactivates a transcriptional regulatory motif. However, in some cases the mutation interferes with regulatory motifs or non-coding RNAs that control the expression of other genes. RNA-seq is clearly the best approach for quickly identifying these situations.

Although mRNA analysis is critical for establishing pathogenicity of deep intronic mutations, biopsy material from affected patient tissues is not always available. This may represent a significant limitation, namely for the study of neurogenetic disorders. Notably, the genes with the longest introns tend to be most highly expressed in the brain (Sibley et al. 2015), and non-canonical splicing mechanisms appear enriched these long introns (Pickrell et al. 2010; Roy and Irimia 2008). Thus, it will be particularly interesting to explore the contribution of deep intronic mutations to human brain disorders and a number of recent possibilities obviate the requirement for brain biopsy. These include using neurons differentiated in vitro from either induced pluripotent stem cells (Bellin et al. 2012) or through direct reprogramming (Tsunemoto et al. 2015).

Acknowledgements We thank Joana Tavares and Isabel Vaz for critical reading of the manuscript. This work was supported by Fundação para a Ciência e a Tecnologia (Grant PTDC/BEX-BCM/5899/2014 and fellowship SFRH/BD/90231/2012 to R.V.D).

References

Abeliovich D, Lavon IP, Lerer I, Cohen T, Springer C, Avital A, Cutting GR (1992) Screening for five mutations detects 97% of cystic fibrosis (CF) chromosomes and predicts a carrier frequency of 1:29 in the Jewish Ashkenazi population. Am J Hum Genet 51:951–956

Akker SA, Misra S, Aslam S, Morgan EL, Smith PJ, Khoo B, Chew SL (2007) Pre-spliceosomal binding of U1 small nuclear ribonucleoprotein (RNP) and heterogenous nuclear RNP E1 is associated with suppression of a growth hormone receptor pseudoexon. Mol Endocrinol 21:2529–2540. doi:10.1210/me.2007-0038

Akman HO et al (2015) Deep intronic GBE1 mutation in manifesting heterozygous patients with adult polyglucosan body disease. JAMA Neurol 72:441–445. doi:10.1001/jamaneurol.2014.4496

Albuisson J et al (2011) Identification of two novel mutations in Shh long-range regulator associated with familial pre-axial polydactyly. Clin Genet 79:371–377. doi:10.1111/j.1399-0004.2010.01465.x

Anczukow O et al (2012) BRCA2 deep intronic mutation causing activation of a cryptic exon: opening toward a new preventive therapeutic strategy. Clin Cancer Res Off J Am Assoc Can Res 18:4903–4909. doi:10.1158/1078-0432.CCR-12-1100

Antonellis A et al (2010) A rare myelin protein zero (MPZ) variant alters enhancer activity in vitro and in vivo. PLoS One 5:e14346. doi:10.1371/journal.pone.0014346

Bagnall RD, Waseem NH, Green PM, Colvin B, Lee C, Giannelli F (1999) Creation of a novel donor splice site in intron 1 of the factor VIII gene leads to activation of a 191 bp cryptic exon in two haemophilia A patients. Br J Haematol 107:766–771

Balz V, Prisack HB, Bier H, Bojar H (2002) Analysis of BRCA1, TP53, and TSG101 germline mutations in German breast and/or ovarian cancer families. Cancer Genet Cytogenet 138:120–127

Barash Y et al (2010) Deciphering the splicing code. Nature 465:53–59. doi:10.1038/nature09000

Baskin B, Gibson WT, Ray PN (2011) Duchenne muscular dystrophy caused by a complex rearrangement between intron 43 of the DMD gene and chromosome 4. Neuromusc Disord NMD 21:178–182. doi:10.1016/j.nmd.2010.11.008

Bauwens M et al (2015) An augmented ABCA4 screen targeting noncoding regions reveals a deep intronic founder variant in Belgian Stargardt patients. Hum Mutat 36:39–42. doi:10.1002/humu.22716

Bax NM et al (2015) Heterozygous deep-intronic variants and deletions in ABCA4 in persons with retinal dystrophies and one exonic ABCA4 variant. Hum Mutat 36:43–47. doi:10.1002/humu.22717

Becker PW et al (2016) An intronic Flk1 enhancer directs arterial-specific expression via RBPJ-mediated venous repression. Arterioscler Thromb Vasc Biol 36:1209–1219. doi:10.1161/ATVBAHA.116.307517

Bellin M, Marchetto MC, Gage FH, Mummery CL (2012) Induced pluripotent stem cells: the new patient? Nat Rev Mol Cell Biol 13:713–726. doi:10.1038/nrm3448

Beltran M et al (2008) A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. Genes Dev 22:756–769. doi:10.1101/gad.455708

Berezikov E, Chung WJ, Willis J, Cuppen E, Lai EC (2007) Mammalian mirtron genes. Mol Cell 28:328–336. doi:10.1016/j.molcel.2007.09.028

Berger A, Maire S, Gaillard MC, Sahel JA, Hantraye P, Bemelmans AP (2016) mRNA trans-splicing in gene therapy for genetic diseases. Wiley Interdiscip Rev RNA 7:487–498. doi:10.1002/wrna.1347

Berglund JA, Chua K, Abovich N, Reed R, Rosbash M (1997) The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. Cell 89:781–787

Bergsma AJ, In 't Groen SL, Verheijen FW, van der Ploeg AT, Pijnappel WP (2016) From cryptic toward canonical pre-mRNA splicing in pompe disease: a pipeline for the development of antisense oligonucleotides molecular therapy. Nucleic Acids 5:e361. doi:10.1038/mtna.2016.75

Berk AJ (2016) Discovery of RNA splicing and genes in pieces. Proc Natl Acad Sci USA 113:801–805. doi:10.1073/pnas.1525084113

Beroud C et al (2004) Dystrophinopathy caused by mid-intronic substitutions activating cryptic exons in the DMD gene. Neuromusc Disord NMD 14:10–18

Bholah Z, Smith MJ, Byers HJ, Miles EK, Evans DG, Newman WG (2014) Intronic splicing mutations in PTCH1 cause Gorlin syndrome. Fam Cancer 13:477–480. doi:10.1007/s10689-014-9712-9

Bieberstein NI, Carrillo Oesterreich F, Straube K, Neugebauer KM (2012) First exon length controls active chromatin signatures and transcription. Cell Rep 2:62–68. doi:10.1016/j.celrep.2012.05.019

Black DL (2003) Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem 72:291–336. doi:10.1146/annurev.biochem.72.121801.161720

Blazquez L et al (2013) In vitro correction of a pseudoexon-generating deep intronic mutation in LGMD2A by antisense oligonucleotides and modified small nuclear RNAs. Hum Mutat 34:1387–1395. doi:10.1002/humu.22379

Bonifert T et al (2014) Pure and syndromic optic atrophy explained by deep intronic OPA1 mutations and an intralocus modifier. Brain J Neurol 137:2164–2177. doi:10.1093/brain/awu165

Bonifert T, Gonzalez Menendez I, Battke F, Theurer Y, Synofzik M, Schols L, Wissinger B (2016) Antisense Oligonucleotide mediated splice correction of a deep intronic mutation in OPA1 molecular therapy. Nucleic Acids 5:e390. doi:10.1038/mtna.2016.93

Bortolin ML, Kiss T (1998) Human U19 intron-encoded snoRNA is processed from a long primary transcript that possesses little potential for protein coding. RNA 4:445–454

Boutz PL, Bhutkar A, Sharp PA (2015) Detained introns are a novel, widespread class of post-transcriptionally spliced introns. Genes Dev 29:63–80. doi:10.1101/gad.247361.114

Bovolenta M et al (2008) A novel custom high density-comparative genomic hybridization array detects common rearrangements as well as deep intronic mutations in dystrophinopathies. BMC Genom 9:572. doi:10.1186/1471-2164-9-572

Bovolenta M et al (2010) Identification of a deep intronic mutation in the COL6A2 gene by a novel custom oligonucleotide CGH array designed to explore allelic and genetic heterogeneity in collagen VI-related myopathies. BMC Med Genet 11:44. doi:10.1186/1471-2350-11-44

Braun TA et al (2013) Non-exomic and synonymous variants in ABCA4 are an important cause of Stargardt disease. Hum Mol Genet 22:5136–5145. doi:10.1093/hmg/ddt367

Brinster RL, Allen JM, Behringer RR, Gelinas RE, Palmiter RD (1988) Introns increase transcriptional efficiency in transgenic mice. Proc Natl Acad Sci USA 85:836–840

Buratti E, Dhir A, Lewandowska MA, Baralle FE (2007) RNA structure is a key regulatory element in pathological ATM and CFTR pseudoexon inclusion events. Nucleic Acids Res 35:4369–4383. doi:10.1093/nar/gkm447

Burnette JM, Miyamoto-Sato E, Schaub MA, Conklin J, Lopez AJ (2005) Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. Genetics 170:661–674. doi:10.1534/genetics.104.039701

Busslinger M, Moschonas N, Flavell RA (1981) Beta+ thalassemia: aberrant splicing results from a single point mutation in an intron. Cell 27:289–298

Caminsky N, Mucaki EJ, Rogan PK (2014) Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. F1000Research 3:282. doi:10.12688/f1000research.5654.1

Camtosun E, Siklar Z, Kocaay P, Ceylaner S, Flanagan SE, Ellard S, Berberoglu M (2015) Three cases of Wolfram syndrome with different clinical aspects. J Pediatr Endocrinol Metab JPEM 28:433–438. doi:10.1515/jpem-2014-0139

Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3:285–298. doi:10.1038/nrg775

Castaman G et al (2011) Deep intronic variations may cause mild hemophilia A. J Thromb Haemost JTH 9:1541–1548. doi:10.1111/j.1538-7836.2011.04408.x

Castellanos E et al (2013) In vitro antisense therapeutics for a deep intronic mutation causing Neurofibromatosis type 2. Eur J Hum Genet EJHG 21:769–773. doi:10.1038/ejhg.2012.261

Chabot B, Shkreta L (2016) Defective control of pre-messenger RNA splicing in human disease. J Cell Biol 212:13–27. doi:10.1083/jcb.201510032

Chen LL (2016) The biogenesis and emerging roles of circular RNAs. Nat Rev Mol Cell Biol 17:205–211. doi:10.1038/nrm.2015.32

Chen M, Manley JL (2009) Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nat Rev Mol Cell Biol 10:741–754. doi:10.1038/nrm2777

Cheng TC et al (1984) beta-Thalassemia in Chinese: use of in vivo RNA analysis and oligonucleotide hybridization in systematic characterization of molecular defects. Proc Natl Acad Sci USA 81:2821–2825

Chillon M et al (1995) A novel donor splice site in intron 11 of the CFTR gene, created by mutation 1811+1.6kbA–>G, produces a new exon: high frequency in Spanish cystic fibrosis chromosomes and association with severe phenotype. Am J Hum Genet 56:623–629

Chmel N et al (2015) A deep-intronic FERMT1 mutation causes kindler syndrome: an explanation for genetically unsolved cases. J Invest Dermatol 135:2876–2879. doi:10.1038/jid.2015.227

Chorev M, Carmel L (2013) Computational identification of functional introns: high positional conservation of introns that harbor RNA genes. Nucleic Acids Res 41:5604–5613. doi:10.1093/nar/gkt244

Clendenning M et al (2011) Mutation deep within an intron of MSH2 causes Lynch syndrome. Fam Cancer 10:297–301. doi:10.1007/s10689-011-9427-0

Corrigan A, Arenas M, Escuredo E, Fairbanks L, Marinaki A (2011) HPRT deficiency: identification of twenty-four novel variants including an unusual deep intronic mutation. Nucleosides Nucleotides Nucleic Acids 30:1260–1265. doi:10.1080/15257770.2011.590172

Corvelo A, Eyras E (2008) Exon creation and establishment in human genes. Genome Biol 9:R141. doi:10.1186/gb-2008-9-9-r141

Costa C et al (2011) A recurrent deep-intronic splicing CF mutation emphasizes the importance of mRNA studies in clinical practice. J Cystic Fibros Off J Eur Cystic Fibros Soc 10:479–482. doi:10.1016/j.jcf.2011.06.011

Costantino L et al (2013) Fine characterization of the recurrent c.1584+18672A>G deep-intronic mutation in the cystic fibrosis transmembrane conductance regulator gene. Am J Respir Cell Mol Biol 48:619–625. doi:10.1165/rcmb.2012-0371OC

Coutinho G, Xie J, Du L, Brusco A, Krainer AR, Gatti RA (2005) Functional significance of a deep intronic mutation in the ATM gene and evidence for an alternative exon 28a. Hum Mutat 25:118–124. doi:10.1002/humu.20170

Cunha KS et al (2016) Hybridization capture-based next-generation sequencing to evaluate coding sequence and deep intronic mutations in the NF1. Gene Genes. doi:10.3390/genes7120133

Damgaard CK, Kahns S, Lykke-Andersen S, Nielsen AL, Jensen TH, Kjems J (2008) A 5′ splice site enhances the recruitment of basal transcription initiation factors in vivo. Mol Cell 29:271–278. doi:10.1016/j.molcel.2007.11.035

David A et al (2007) An intronic growth hormone receptor mutation causing activation of a pseudoexon is associated with a broad spectrum of growth hormone insensitivity phenotypes. J Clin Endocrinol Metab 92:655–659. doi:10.1210/jc.2006-1527

Davis RL, Homer VM, George PM, Brennan SO (2009) A deep intronic mutation in FGB creates a consensus exonic splicing enhancer motif that results in afibrinogenemia caused by aberrant mRNA splicing, which can be corrected in vitro with antisense oligonucleotide treatment. Hum Mutat 30:221–227. doi:10.1002/humu.20839

De Klein A et al (1998) A G–>A transition creates a branch point sequence and activation of a cryptic exon, resulting in the hereditary disorder neurofibromatosis 2. Hum Mol Genet 7:393–398

Deburgrave N et al (2007) Protein- and mRNA-based phenotype-genotype correlations in DMD/BMD with point mutations and molecular basis for BMD with nonsense and frameshift mutations in the DMD gene. Hum Mutat 28:183–195. doi:10.1002/humu.20422

Dehainault C et al (2007) A deep intronic mutation in the RB1 gene leads to intronic sequence exonisation. Eur J Hum Genet EJHG 15:473–477. doi:10.1038/sj.ejhg.5201787

den Hollander AI et al (2006) Mutations in the CEP290 (NPHP6) gene are a frequent cause of Leber congenital amaurosis. Am J Hum Genet 79:556–561. doi:10.1086/507318

Dhir A, Buratti E (2010) Alternative splicing: role of pseudoexons in human disease and potential therapeutic strategies. FEBS J 277:841–855. doi:10.1111/j.1742-4658.2009.07520.x

Dobkin C, Pergolizzi RG, Bahre P, Bank A (1983) Abnormal splice in a mutant human beta-globin gene not at the site of a mutation. Proc Natl Acad Sci USA 80:1184–1188

Dominov JA, Uyan O, Sapp PC, McKenna-Yasek D, Nallamilli BR, Hegde M, Brown RH Jr (2014) A novel dysferlin mutant pseudoexon bypassed with antisense oligonucleotides. Ann Clin Transl Neurol 1:703–720. doi:10.1002/acn3.96

Dong XY et al (2008) SnoRNA U50 is a candidate tumor-suppressor gene at 6q14.3 with a mutation associated with clinically significant prostate cancer. Hum Mol Genet 17:1031–1042. doi:10.1093/hmg/ddm375

Dong XY, Guo P, Boyd J, Sun X, Li Q, Zhou W, Dong JT (2009) Implication of snoRNA U50 in human breast cancer. J Genet Genom 36:447–454. doi:10.1016/S1673-8527(08)60134-4

Dong R, Ma XK, Chen LL, Yang L (2016) Increased complexity of circRNA expression during species evolution. RNA Biol. doi:10.1080/15476286.2016.1269999

Dreyfuss G, Kim VN, Kataoka N (2002) Messenger-RNA-binding proteins and the messages they carry. Nat Rev Mol Cell Biol 3:195–205. doi:10.1038/nrm760

Duff MO et al (2015) Genome-wide identification of zero nucleotide recursive splicing in Drosophila. Nature 521:376–379. doi:10.1038/nature14475

Edery P et al (2011) Association of TALS developmental disorder with defect in minor splicing component U4atac snRNA. Science 332:240–243. doi:10.1126/science.1202205

Faa V et al (2009) Characterization of a disease-associated mutation affecting a putative splicing regulatory element in intron 6b of the cystic fibrosis transmembrane conductance regulator (CFTR) gene. J Biol Chem 284:30024–30031. doi:10.1074/jbc.M109.032623

Ferlini A, Galie N, Merlini L, Sewry C, Branzi A, Muntoni F (1998) A novel Alu-like element rearranged in the dystrophin gene causes a splicing mutation in a family with X-linked dilated cardiomyopathy. Am J Hum Genet 63:436–446. doi:10.1086/301952

Ferri L, Covello G, Caciotti A, Guerrini R, Denti MA, Morrone A (2016) Double-target antisense U1snRNAs correct mis-splicing due to c.639+861C>T and c.639+919G>A GLA deep intronic mutations molecular therapy. Nucleic Acids 5:380. doi:10.1038/mtna.2016.88

Flanagan SE et al (2013) Next-generation sequencing reveals deep intronic cryptic ABCC8 and HADH splicing founder mutations causing hyperinsulinism by pseudoexon activation. Am J Hum Genet 92:131–136. doi:10.1016/j.ajhg.2012.11.017

Flynt AS, Greimann JC, Chung WJ, Lima CD, Lai EC (2010) MicroRNA biogenesis via splicing and exosome-mediated trimming in Drosophila. Mol Cell 38:900–907. doi:10.1016/j.molcel.2010.06.014

Friedman KJ, Kole J, Cohn JA, Knowles MR, Silverman LM, Kole R (1999) Correction of aberrant splicing of the cystic fibrosis transmembrane conductance regulator (CFTR) gene by antisense oligonucleotides. J Biol Chem 274:36193–36199

Friedrich K et al (2010) WRN mutations in Werner syndrome patients: genomic rearrangements, unusual intronic mutations and ethnic-specific alterations. Hum Genet 128:103–111. doi:10.1007/s00439-010-0832-5

Furniss D, Lettice LA, Taylor IB, Critchley PS, Giele H, Hill RE, Wilkie AO (2008) A variant in the sonic hedgehog regulatory sequence (ZRS) is associated with triphalangeal thumb and

deregulates expression in the developing limb. Hum Mol Genet 17:2417–2423. doi:10.1093/hmg/ddn141
Gaffney DJ, Keightley PD (2004) Unexpected conserved non-coding DNA blocks in mammals. Trends Genet TIG 20:332–337. doi:10.1016/j.tig.2004.06.011
Gallagher RC, Pils B, Albalwi M, Francke U (2002) Evidence for the role of PWCR1/HBII-85 C/D box small nucleolar RNAs in Prader–Willi syndrome. Am J Hum Genet 71:669–678. doi:10.1086/342408
Gazzoli I, Pulyakhina I, Verwey NE, Ariyurek Y, Laros JF, t Hoen PA, Aartsma-Rus A (2016) Non-sequential and multi-step splicing of the dystrophin transcript. RNA Biol 13:290–305. doi:10.1080/15476286.2015.1125074
Gilbert W (1978) Why genes in pieces? Nature 271:501
Gilissen C et al (2014) Genome sequencing identifies major causes of severe intellectual disability. Nature 511:344–347. doi:10.1038/nature13394
Gillis E et al (2014) An FBN1 deep intronic mutation in a familial case of Marfan syndrome: an explanation for genetically unsolved cases? Hum Mutat 35:571–574. doi:10.1002/humu.22540
Gonorazky H et al (2016) RNAseq analysis for the diagnosis of muscular dystrophy. Ann Clin Transl Neurol 3:55–60. doi:10.1002/acn3.267
Greer K et al (2015) Pseudoexon activation increases phenotype severity in a Becker muscular dystrophy patient. Mol Genet Genom Med 3:320–326. doi:10.1002/mgg3.144
Gudipati RK, Xu Z, Lebreton A, Seraphin B, Steinmetz LM, Jacquier A, Libri D (2012) Extensive degradation of RNA precursors by the exosome in wild-type cells. Mol Cell 48:409–421. doi:10.1016/j.molcel.2012.08.018
Guo DC, Gupta P, Tran-Fadulu V, Guidry TV, Leduc MS, Schaefer FV, Milewicz DM (2008) An FBN1 pseudoexon mutation in a patient with Marfan syndrome: confirmation of cryptic mutations leading to disease. J Hum Genet 53:1007–1011. doi:10.1007/s10038-008-0334-7
Gurnett CA, Bowcock AM, Dietz FR, Morcuende JA, Murray JC, Dobbs MB (2007) Two novel point mutations in the long-range SHH enhancer in three families with triphalangeal thumb and preaxial polydactyly. Am J Med Genet Part A 143A:27–32. doi:10.1002/ajmg.a.31563
Gurvich OL et al (2008) DMD pseudoexon mutations: splicing efficiency, phenotype, and potential therapy. Ann Neurol 63:81–89. doi:10.1002/ana.21290
Hall SL, Padgett RA (1996) Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns. Science 271:1716–1718
Hang J, Wan R, Yan C, Shi Y (2015) Structural basis of pre-mRNA splicing. Science 349:1191–1198. doi:10.1126/science.aac8159
Hare MP, Palumbi SR (2003) High intron sequence conservation across three mammalian orders suggests functional constraints. Mol Biol Evol 20:969–978. doi:10.1093/molbev/msg111
Harland M, Mistry S, Bishop DT, Bishop JA (2001) A deep intronic mutation in CDKN2A is associated with disease in a subset of melanoma pedigrees. Hum Mol Genet 10:2679–2686
Hatton AR, Subramaniam V, Lopez AJ (1998) Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon–exon junctions. Mol Cell 2:787–796
He H et al (2011) Mutations in U4atac snRNA, a component of the minor spliceosome, in the developmental disorder MOPD I. Science 332:238–240. doi:10.1126/science.1200587
Heinzen EL et al (2008) Tissue-specific genetic control of splicing: implications for the study of complex traits. PLoS Biol 6:e1. doi:10.1371/journal.pbio.1000001
Highsmith WE et al (1994) A novel mutation in the cystic fibrosis gene in patients with pulmonary disease but normal sweat chloride concentrations. N Engl J Med 331:974–980. doi:10.1056/NEJM199410133311503
Hilgert N, Topsakal V, van Dinther J, Offeciers E, Van de Heyning P, Van Camp G (2008) A splice-site mutation and overexpression of MYO6 cause a similar phenotype in two families with autosomal dominant hearing loss. Eur J Hum Genet EJHG 16:593–602. doi:10.1038/sj.ejhg.5202000
Homolova K, Zavadakova P, Doktor TK, Schroeder LD, Kozich V, Andresen BS (2010) The deep intronic c.903+469T>C mutation in the MTRR gene creates an SF2/ASF binding exonic splicing enhancer, which leads to pseudoexon activation and causes the cblE type of homocystinuria. Hum Mutat 31:437–444. doi:10.1002/humu.21206
Hsiao YH, Bahn JH, Lin X, Chan TM, Wang R, Xiao X (2016) Alternative splicing modulated by genetic variants demonstrates accelerated evolution regulated by highly conserved proteins. Genome Res 26:440–450. doi:10.1101/gr.193359.115
Hube F, Francastel C (2015) Mammalian introns: when the junk generates molecular diversity. Int J Mol Sci 16:4429–4452. doi:10.3390/ijms16034429
Ikeda H et al (1997) Molecular analysis of dihydropteridine reductase deficiency: identification of two novel mutations in Japanese patients. Hum Genet 100:637–642
Ikezawa M, Nishino I, Goto Y, Miike T, Nonaka I (1999) Newly recognized exons induced by a splicing abnormality from an intronic mutation of the dystrophin gene resulting in Duchenne muscular dystrophy. Mutations in brief no. 213. Online. Hum Mutat 13:170. doi:10.1002/(SICI)1098-1004(1999)13:2<170:AID-HUMU12>3.0.CO;2-7
Inaba H, Koyama T, Shinozawa K, Amano K, Fukutake K (2013) Identification and characterization of an adenine to guanine transition within intron 10 of the factor VIII gene as a causative mutation in a patient with mild haemophilia A. Haemoph Off J World Feder Hemoph 19:100–105. doi:10.1111/j.1365-2516.2012.02906.x
Irimia M, Roy SW (2014) Origin of spliceosomal introns and alternative splicing. Cold Spring Harbor Perspect Biol. doi:10.1101/cshperspect.a016071
Ishii S, Nakao S, Minamikawa-Tachino R, Desnick RJ, Fan JQ (2002) Alternative splicing in the alpha-galactosidase A gene: increased exon inclusion results in the Fabry cardiac phenotype. Am J Hum Genet 70:994–1002. doi:10.1086/339431
Jafarifar F, Dietrich RC, Hiznay JM, Padgett RA (2014) Biochemical defects in minor spliceosome function in the developmental disorder MOPD I. RNA 20:1078–1089. doi:10.1261/rna.045187.114
Jaillon O et al (2008) Translational control of intron splicing in eukaryotes. Nature 451:359–362. doi:10.1038/nature06495
Janz S, Potter M, Rabkin CS (2003) Lymphoma- and leukemia-associated chromosomal translocations in healthy individuals. Genes Chromosom Cancer 36:211–223. doi:10.1002/gcc.10178
Jeck WR et al (2013) Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA 19:141–157. doi:10.1261/rna.035667.112
Juneau K, Miranda M, Hillenmeyer ME, Nislow C, Davis RW (2006) Introns regulate RNA and protein abundance in yeast. Genetics 174:511–518. doi:10.1534/genetics.106.058560
Kaida D, Berg MG, Younis I, Kasim M, Singh LN, Wan L, Dreyfuss G (2010) U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature 468:664–668. doi:10.1038/nature09479
Kansakoski J et al (2016) Complete androgen insensitivity syndrome caused by a deep intronic pseudoexon-activating mutation in the androgen receptor gene. Sci Rep 6:32819. doi:10.1038/srep32819
Kelly S et al (2015) Splicing of many human genes involves sites embedded within introns. Nucleic Acids Res 43:4721–4732. doi:10.1093/nar/gkv386

Keren H, Lev-Maor G, Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet 11:345–355. doi:10.1038/nrg2776

Khelifi MM et al (2011) Pure intronic rearrangements leading to aberrant pseudoexon inclusion in dystrophinopathy: a new class of mutations? Hum Mutat 32:467–475. doi:10.1002/humu.21471

King K, Flinter FA, Nihalani V, Green PM (2002) Unusual deep intronic mutations in the COL4A5 gene cause X linked Alport syndrome. Hum Genet 111:548–554. doi:10.1007/s00439-002-0830-3

Knebelmann B et al (1995) Splice-mediated insertion of an Alu sequence in the COL4A3 mRNA causing autosomal recessive Alport syndrome. Hum Mol Genet 4:675–679

Kollberg G et al (2009) Clinical manifestation and a new ISCU mutation in iron-sulphur cluster deficiency myopathy. Brain J Neurol 132:2170–2179. doi:10.1093/brain/awp152

Konarska MM, Padgett RA, Sharp PA (1985) Trans splicing of mRNA precursors in vitro. Cell 42:165–171

Krawczak M, Thomas NS, Hundrieser B, Mort M, Wittig M, Hampe J, Cooper DN (2007) Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. Hum Mutat 28:150–158. doi:10.1002/humu.20400

Kurio H, Murayama E, Kaneko T, Shibata Y, Inai T, Iida H (2008) Intron retention generates a novel isoform of CEACAM6 that may act as an adhesion molecule in the ectoplasmic specialization structures between spermatids and sertoli cells in rat testis. Biol Reprod 79:1062–1073. doi:10.1095/biolreprod.108.069872

Kwek KY et al (2002) U1 snRNA associates with TFIIH and regulates transcriptional initiation. Nat Struct Biol 9:800–805. doi:10.1038/nsb862

Lebon S et al (2007) A novel mutation of the NDUFS7 gene leads to activation of a cryptic exon and impaired assembly of mitochondrial complex I in a patient with Leigh syndrome. Mol Genet Metab 92:104–108. doi:10.1016/j.ymgme.2007.05.010

Lee WI, Torgerson TR, Schumacher MJ, Yel L, Zhu Q, Ochs HD (2005) Molecular analysis of a large cohort of patients with the hyper immunoglobulin M (IgM) syndrome. Blood 105:1881–1890. doi:10.1182/blood-2003-12-4420

Lettice LA et al (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet 12:1725–1735

Li H, Wang J, Mor G, Sklar J (2008) A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. Science 321:1357–1361. doi:10.1126/science.1156725

Liang D, Wilusz JE (2014) Short intronic repeat sequences facilitate circular RNA production. Genes Develop 28:2233–2247. doi:10.1101/gad.251926.114

Liquori A et al (2016) Whole USH2A gene sequencing identifies several new deep intronic mutations. Hum Mutat 37:184–193. doi:10.1002/humu.22926

Liu HX, Zhang M, Krainer AR (1998) Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev 12:1998–2012

Liu Z et al (2001) Structural basis for recognition of the intron branch site RNA by splicing factor 1. Science 294:1098–1102. doi:10.1126/science.1064719

Lo YF et al (2011) Recurrent deep intronic mutations in the SLC12A3 gene responsible for Gitelman's syndrome. Clin J Am Soc Nephrol CJASN 6:630–639. doi:10.2215/CJN.06730810

Long M, Betran E, Thornton K, Wang W (2003) The origin of new genes: glimpses from the young and old. Nat Rev Genet 4:865–875. doi:10.1038/nrg1204

Lopez-Bigas N, Audit B, Ouzounis C, Parra G, Guigo R (2005) Are splicing mutations the most frequent cause of hereditary disease? FEBS Lett 579:1900–1903. doi:10.1016/j.febslet.2005.02.047

Madden HR, Fletcher S, Davis MR, Wilton SD (2009) Characterization of a complex Duchenne muscular dystrophy-causing dystrophin gene inversion and restoration of the reading frame by induced exon skipping. Hum Mutat 30:22–28. doi:10.1002/humu.20806

Mancini C et al (2012) Megalencephalic leukoencephalopathy with subcortical cysts type 1 (MLC1) due to a homozygous deep intronic splicing mutation (c.895-226T>G) abrogated in vitro using an antisense morpholino oligonucleotide. Neurogenetics 13:205–214. doi:10.1007/s10048-012-0331-z

Masala MV et al (2007) Epidemiology and clinical aspects of Werner's syndrome in North Sardinia: description of a cluster. Eur J Dermatol EJD 17:213–216. doi:10.1684/ejd.2007.0155

Mattick JS (2001) Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep 2:986–991. doi:10.1093/embo-reports/kve230

Mattick JS, Gagen MJ (2001) The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. Mol Biol Evol 18:1611–1630

Mauger O, Lemoine F, Scheiffele P (2016) Targeted intron retention and excision for rapid gene regulation in response to neuronal activity. Neuron 92:1266–1278. doi:10.1016/j.neuron.2016.11.032

Maxwell ES, Fournier MJ (1995) The small nucleolar RNAs. Annu Rev Biochem 64:897–934. doi:10.1146/annurev.bi.64.070195.004341

Mayer K, Ballhausen W, Leistner W, Rott H (2000) Three novel types of splicing aberrations in the tuberous sclerosis TSC2 gene caused by mutations apart from splice consensus sequences. Biochem Biophys Acta 1502:495–507

McConville CM, Stankovic T, Byrd PJ, McGuire GM, Yao QY, Lennox GG, Taylor MR (1996) Mutations associated with variant phenotypes in ataxia-telangiectasia. Am J Hum Genet 59:320–330

McKenzie RW, Brennan MD (1996) The two small introns of the Drosophila affinidisjuncta Adh gene are required for normal transcription. Nucleic Acids Res 24:3635–3642

Merendino L, Guth S, Bilbao D, Martinez C, Valcarcel J (1999) Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF35 and the 3′ splice site AG. Nature 402:838–841. doi:10.1038/45602

Merico D et al (2015) Compound heterozygous mutations in the noncoding RNU4ATAC cause Roifman Syndrome by disrupting minor intron splicing. Nature Commun 6:8718. doi:10.1038/ncomms9718

Mertens F, Johansson B, Fioretos T, Mitelman F (2015) The emerging complexity of gene fusions in cancer. Nat Rev Cancer 15:371–381. doi:10.1038/nrc3947

Metherell LA et al (2001) Pseudoexon activation as a novel mechanism for disease resulting in atypical growth-hormone insensitivity. Am J Hum Genet 69:641–646. doi:10.1086/323266

Michel-Calemard L et al (2009) Pseudoexon activation in the PKHD1 gene: a French founder intronic mutation IVS46+653A>G causing severe autosomal recessive polycystic kidney disease. Clin Genet 75:203–206. doi:10.1111/j.1399-0004.2008.01106.x

Mitchell GA et al (1991) Splice-mediated insertion of an Alu sequence inactivates ornithine delta-aminotransferase: a role for Alu elements in human mutation. Proc Natl Acad Sci USA 88:815–819

Mochel F et al (2008) Splice mutation in the iron-sulfur cluster scaffold protein ISCU causes myopathy with exercise intolerance. Am J Hum Genet 82:652–660. doi:10.1016/j.ajhg.2007.12.012

Monnier N, Gout JP, Pin I, Gauthier G, Lunardi J (2001) A novel 3600+11.5 kb C>G homozygous splicing mutation in a black African, consanguineous CF family. J Med Genet 38:E4

Monnier N, Ferreiro A, Marty I, Labarre-Vila A, Mezin P, Lunardi J (2003) A homozygous splicing mutation causing a depletion of skeletal muscle RYR1 is associated with multi-minicore disease congenital myopathy with ophthalmoplegia. Hum Mol Genet 12:1171–1178

Naftelberg S, Schor IE, Ast G, Kornblihtt AR (2015) Regulation of alternative splicing through coupling with transcription and chromatin structure. Annu Rev Biochem 84:165–198. doi:10.1146/annurev-biochem-060614-034242

Naro C et al (2017) An orchestrated intron retention program in meiosis controls timely usage of transcripts during germ cell differentiation. Dev Cell 41(82–93):e84. doi:10.1016/j.devcel.2017.03.003

Naruto T, Okamoto N, Masuda K, Endo T, Hatsukawa Y, Kohmoto T, Imoto I (2015) Deep intronic GPR143 mutation in a Japanese family with ocular albinism. Sci Rep 5:11334. doi:10.1038/srep11334

Nathan N, Girodon E, Clement A, Corvol H (2012) A rare CFTR intronic mutation related to a mild CF disease in a 12-year-old girl. BMJ Case Rep. doi:10.1136/bcr-2012-006918

Noack D, Heyworth PG, Newburger PE, Cross AR (2001) An unusual intronic mutation in the CYBB gene giving rise to chronic granulomatous disease. Biochem Biophys Acta 1537:125–131

Nozu K et al (2009) A deep intronic mutation in the SLC12A3 gene leads to Gitelman syndrome. Pediatr Res 66:590–593. doi:10.1203/PDR.0b013e3181b9b4d3

Ogino W et al (2007) Mutation analysis of the ornithine transcarbamylase (OTC) gene in five Japanese OTC deficiency patients revealed two known and three novel mutations including a deep intronic mutation. Kobe J Med Sci 53:229–240

Okamura K, Hagen JW, Duan H, Tyler DM, Lai EC (2007) The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. Cell 130:89–100. doi:10.1016/j.cell.2007.06.028

Olsson A, Lind L, Thornell LE, Holmberg M (2008) Myopathy with lactic acidosis is linked to chromosome 12q23.3-24.11 and caused by an intron mutation in the ISCU gene resulting in a splicing defect. Hum Mol Genet 17:1666–1672. doi:10.1093/hmg/ddn057

Padgett RA (2012) New connections between splicing and human disease. Trends Genet TIG 28:147–154. doi:10.1016/j.tig.2012.01.001

Pagani F, Buratti E, Stuani C, Bendix R, Dork T, Baralle FE (2002) A new type of mutation causes a splicing defect in ATM. Nat Genet 30:426–429. doi:10.1038/ng858

Palagano E et al (2015) Buried in the Middle but guilty: intronic mutations in the TCIRG1 gene cause human autosomal recessive osteopetrosis. J Bone Miner Res Off J Am Soc Bone Miner Res 30:1814–1821. doi:10.1002/jbmr.2517

Palhais B, Dembic M, Sabaratnam R, Nielsen KS, Doktor TK, Bruun GH, Andresen BS (2016) The prevalent deep intronic c. 639+919 G>A GLA mutation causes pseudoexon activation and Fabry disease by abolishing the binding of hnRNPA1 and hnRNP A2/B1 to a splicing silencer. Mol Genet Metab 119:258–269. doi:10.1016/j.ymgme.2016.08.007

Papasaikas P, Valcarcel J (2016) The spliceosome: the ultimate RNA chaperone and sculptor. Trends Biochem Sci 41:33–45. doi:10.1016/j.tibs.2015.11.003

Park SG, Hannenhalli S, Choi SS (2014) Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals. BMC Genom 15:526. doi:10.1186/1471-2164-15-526

Patel AA, Steitz JA (2003) Splicing double: insights from the second spliceosome. Nat Rev Mol Cell Biol 4:960–970. doi:10.1038/nrm1259

Pedrotti S, Cooper TA (2014) In Brief: (mis)splicing in disease. J Pathol 233:1–3. doi:10.1002/path.4337

Pezeshkpoor B et al (2013) Deep intronic 'mutations' cause hemophilia A: application of next generation sequencing in patients without detectable mutation in F8 cDNA. J Thromb Haemost JTH 11:1679–1687. doi:10.1111/jth.12339

Pickrell JK, Pai AA, Gilad Y, Pritchard JK (2010) Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genet 6:e1001236. doi:10.1371/journal.pgen.1001236

Plate M, Duga S, Castaman G, Rodeghiero F, Asselta R (2009) Recurrence of the 'deep-intronic' FGG IVS6-320A>T mutation causing quantitative fibrinogen deficiency in the Italian population of Veneto. Blood Coagul Fibrinolysis Int J Haemost Thromb 20:381–384

Popp MW, Maquat LE (2013) Organizing principles of mammalian nonsense-mediated mRNA decay. Annu Rev Genet 47:139–165. doi:10.1146/annurev-genet-111212-133424

Purevsuren J, Fukao T, Hasegawa Y, Fukuda S, Kobayashi H, Yamaguchi S (2008) Study of deep intronic sequence exonization in a Japanese neonate with a mitochondrial trifunctional protein deficiency. Mol Genet Metab 95:46–51. doi:10.1016/j.ymgme.2008.06.013

Puttaraju M, Jamison SF, Mansfield SG, Garcia-Blanco MA, Mitchell LG (1999) Spliceosome-mediated RNA trans-splicing as a tool for gene therapy. Nat Biotechnol 17:246–252. doi:10.1038/6986

Rathmann M, Bunge S, Beck M, Kresse H, Tylki-Szymanska A, Gal A (1996) Mucopolysaccharidosis type II (Hunter syndrome): mutation "hot spots" in the iduronate-2-sulfatase gene. Am J Hum Genet 59:1202–1209

Richards AJ, McNinch A, Whittaker J, Treacy B, Oakhill K, Poulson A, Snead MP (2012) Splicing analysis of unclassified variants in COL2A1 and COL11A1 identifies deep intronic pathogenic mutations. Eur J Hum Genet 20:552–558. doi:10.1038/ejhg.2011.223

Rickman DS et al (2009) SLC45A3-ELK4 is a novel and frequent erythroblast transformation-specific fusion transcript in prostate cancer. Can Res 69:2734–2738. doi:10.1158/0008-5472.CAN-08-4926

Rincon A, Aguado C, Desviat LR, Sanchez-Alcudia R, Ugarte M, Perez B (2007) Propionic and methylmalonic acidemia: antisense therapeutics for intronic variations causing aberrantly spliced messenger RNA. Am J Hum Genet 81:1262–1270. doi:10.1086/522376

Rio Frio T, McGee TL, Wade NM, Iseli C, Beckmann JS, Berson EL, Rivolta C (2009) A single-base substitution within an intronic repetitive element causes dominant retinitis pigmentosa with reduced penetrance. Hum Mutat 30:1340–1347. doi:10.1002/humu.21071

Roca X, Sachidanandam R, Krainer AR (2003) Intrinsic differences between authentic and cryptic 5′ splice sites. Nucleic Acids Res 31:6321–6333

Roca X, Akerman M, Gaus H, Berdeja A, Bennett CF, Krainer AR (2012) Widespread recognition of 5′ splice sites by noncanonical base-pairing to U1 snRNA involving bulged nucleotides. Genes Dev 26:1098–1109. doi:10.1101/gad.190173.112

Roca X, Krainer AR, Eperon IC (2013) Pick one, but be quick: 5′ splice sites and the problems of too many choices. Genes Dev 27:129–144. doi:10.1101/gad.209759.112

Rodriguez-Pascau L, Coll MJ, Vilageliu L, Grinberg D (2009) Antisense oligonucleotide treatment for a pseudoexon-generating mutation in the NPC1 gene causing Niemann-Pick type C disease. Hum Mutat 30:E993–E1001. doi:10.1002/humu.21119

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr Biol 13:1512–1517

Romano M, Buratti E, Baralle D (2013) Role of pseudoexons and pseudointrons in human cancer. Int J Cell Biol 2013:810572. doi:10.1155/2013/810572

Roy SW, Irimia M (2008) Intron mis-splicing: no alternative? Genome Biol 9:208. doi:10.1186/gb-2008-9-2-208

Ruan GX, Barry E, Yu D, Lukason M, Cheng SH, Scaria A (2017) CRISPR/Cas9-mediated genome editing as a therapeutic approach for leber congenital amaurosis 10. Mol Ther J Am Soc Gene Ther 25:331–341. doi:10.1016/j.ymthe.2016.12.006

Rump A et al (2006) A splice-supporting intronic mutation in the last bp position of a cryptic exon within intron 6 of the CYBB gene induces its incorporation into the mRNA causing chronic granulomatous disease (CGD). Gene 371:174–181. doi:10.1016/j.gene.2005.11.036

Runte M, Huttenhofer A, Gross S, Kiefmann M, Horsthemke B, Buiting K (2001) The IC-SNURF-SNRPN transcript serves as a host for multiple small nucleolar RNA species and as an antisense RNA for UBE3A. Hum Mol Genet 10:2687–2700

Ruskin B, Green MR (1985) An RNA processing activity that debranches RNA lariats. Science 229:135–140

Sagai T, Hosoya M, Mizushina Y, Tamura M, Shiroishi T (2005) Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. Development 132:797–803. doi:10.1242/dev.01613

Sahoo T et al (2008) Prader–Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. Nat Genet 40:719–721. doi:10.1038/ng.158

Sakamoto O, Ohura T, Katsushima Y, Fujiwara I, Ogawa E, Miyabayashi S, Iinuma K (2001) A novel intronic mutation of the TAZ (G4.5) gene in a patient with Barth syndrome: creation of a 5′ splice donor site with variant GC consensus and elongation of the upstream exon. Hum Genet 109:559–563. doi:10.1007/s00439-001-0612-3

Salzman J (2016) Circular RNA expression: its potential regulation and function trends in genetics. TIG 32:309–316. doi:10.1016/j.tig.2016.03.002

Schlackow M, Nojima T, Gomes T, Dhir A, Carmo-Fonseca M, Proudfoot NJ (2017) Distinctive patterns of transcription and RNA processing for human lincRNAs. Mol Cell 65:25–38. doi:10.1016/j.molcel.2016.11.029

Schneider A, Maas SM, Hennekam RC, Hanauer A (2013) Identification of the first deep intronic mutation in the RPS6KA3 gene in a patient with a severe form of Coffin-Lowry syndrome. Eur J Med Genet 56:150–152. doi:10.1016/j.ejmg.2012.11.007

Schollen E et al (2007) Characterization of two unusual truncating PMM2 mutations in two CDG-Ia patients. Mol Genet Metab 90:408–413. doi:10.1016/j.ymgme.2007.01.003

Schulz HL et al (2017) Mutation spectrum of the ABCA4 gene in 335 stargardt disease patients from a multicenter german cohort-impact of selected deep intronic variants and common SNPs. Invest Ophthalmol Vis Sci 58:394–403. doi:10.1167/iovs.16-19936

Scotti MM, Swanson MS (2016) RNA mis-splicing in disease. Nat Rev Genet 17:19–32. doi:10.1038/nrg.2015.3

Seraphin B, Rosbash M (1989) Identification of functional U1 snRNA-pre-mRNA complexes committed to spliceosome assembly and splicing. Cell 59:349–358

Shabalina SA, Ogurtsov AY, Spiridonov AN, Novichkov PS, Spiridonov NA, Koonin EV (2010) Distinct patterns of expression and evolution of intronless and intron-containing mammalian genes. Mol Biol Evol 27:1745–1749. doi:10.1093/molbev/msq086

Sharp PA, Konarksa MM, Grabowski PJ, Lamond AI, Marciniak R, Seiler SR (1987) Splicing of messenger RNA precursors. Cold Spring Harb Symp Quant Biol 52:277–285

Sibley CR et al (2015) Recursive splicing in long vertebrate genes. Nature 521:371–375. doi:10.1038/nature14466

Sibley CR, Blazquez L, Ule J (2016) Lessons from non-canonical splicing. Nat Rev Genet 17:407–421. doi:10.1038/nrg.2016.46

Singh RK, Cooper TA (2012) Pre-mRNA splicing in disease and therapeutics. Trends Mol Med 18:472–482. doi:10.1016/j.molmed.2012.06.006

Siprashvili Z et al (2016) The noncoding RNAs SNORD50A and SNORD50B bind K-Ras and are recurrently deleted in human cancer. Nat Genet 48:53–58. doi:10.1038/ng.3452

Sironi M et al (2004) Silencer elements as possible inhibitors of pseudoexon splicing. Nucleic Acids Res 32:1783–1791. doi:10.1093/nar/gkh341

Solnick D (1985) Trans splicing of mRNA precursors. Cell 42:157–164

Spena S, Asselta R, Plate M, Castaman G, Duga S, Tenchini ML (2007) Pseudo-exon activation caused by a deep-intronic mutation in the fibrinogen gamma-chain gene as a novel mechanism for congenital afibrinogenaemia. Br J Haematol 139:128–132. doi:10.1111/j.1365-2141.2007.06758.x

Spier I et al (2012) Deep intronic APC mutations explain a substantial proportion of patients with familial or early-onset adenomatous polyposis. Hum Mutat 33:1045–1050. doi:10.1002/humu.22082

Spritz RA et al (1981) Base substitution in an intervening sequence of a beta+-thalassemic human globin gene. Proc Natl Acad Sci USA 78:2455–2459

Stadhouders R, van den Heuvel A, Kolovos P, Jorna R, Leslie K, Grosveld F, Soler E (2012) Transcription regulation by distal enhancers: who's in the loop? Transcription 3:181–186. doi:10.4161/trns.20720

Stenson PD, Ball EV, Mort M, Phillips AD, Shaw K, Cooper DN (2012) The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. Curr Protoc Bioinf Chapter 1(Unit1):13. doi:10.1002/0471250953.bi0113s39

Sterne-Weiler T, Sanford JR (2014) Exon identity crisis: disease-causing mutations that disrupt the splicing code. Genome Biol 15:201. doi:10.1186/gb4150

Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR (2011) Loss of exon identity is a common mechanism of human inherited disease. Genome Res 21:1563–1571. doi:10.1101/gr.118638.110

Straniero L et al (2016) Whole-gene CFTR sequencing combined with digital RT-PCR improves genetic diagnosis of cystic fibrosis. J Hum Genet 61:977–984. doi:10.1038/jhg.2016.101

Stum M et al (2006) Spectrum of HSPG2 (Perlecan) mutations in patients with Schwartz–Jampel syndrome. Hum Mutat 27:1082–1091. doi:10.1002/humu.20388

Svaasand EK, Engebretsen LF, Ludvigsen T, Brechan W, Sjursen W (2015) A novel deep intronic mutation introducing a cryptic exon causing neurofibromatosis type 1 in a family with highly variable phenotypes: a case study 4. doi:10.4172/2161-1041.1000152

Szafranski P, Yang Y, Nelson MU, Bizzarro MJ, Morotti RA, Langston C, Stankiewicz P (2013) Novel FOXF1 deep intronic deletion causes lethal lung developmental disorder, alveolar capillary dysplasia with misalignment of pulmonary veins. Hum Mutat 34:1467–1471. doi:10.1002/humu.22395

Takeshima Y et al (2010) Mutation spectrum of the dystrophin gene in 442 Duchenne/Becker muscular dystrophy cases from one Japanese referral center. J Hum Genet 55:379–388. doi:10.1038/jhg.2010.49

Tarn WY, Steitz JA (1996a) Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT–AC introns. Science 273:1824–1832

Tarn WY, Steitz JA (1996b) A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT–AC) intron in vitro. Cell 84:801–811

Trabelsi M, Beugnet C, Deburgrave N, Commere V, Orhant L, Leturcq F, Chelly J (2014) When a mid-intronic variation of DMD gene creates an ESE site. Neuromusc Disord 24:1111–1117. doi:10.1016/j.nmd.2014.07.003

Treisman R, Orkin SH, Maniatis T (1983) Specific transcription and RNA splicing defects in five cloned beta-thalassaemia genes. Nature 302:591–596

Tsunemoto RK, Eade KT, Blanchard JW, Baldwin KK (2015) Forward engineering neuronal diversity using direct reprogramming. EMBO J 34:1445–1455. doi:10.15252/embj.201591402

Tsuruta M et al (1998) Molecular basis of intermittent maple syrup urine disease: novel mutations in the E2 gene of the branched-chain alpha-keto acid dehydrogenase complex. J Hum Genet 43:91–100. doi:10.1007/s100380050047

Tuffery-Giraud S, Saquet C, Chambert S, Claustres M (2003) Pseudoexon activation in the DMD gene as a novel mechanism for Becker muscular dystrophy. Hum Mutat 21:608–614. doi:10.1002/humu.10214

Tycowski KT, Shu MD, Steitz JA (1996) A mammalian gene with introns instead of exons generating stable RNA products. Nature 379:464–466. doi:10.1038/379464a0

Valdmanis PN et al (2009) A mutation that creates a pseudoexon in SOD1 causes familial ALS. Ann Hum Genet 73:652–657

Valen E et al (2011) Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes. Nat Struct Mol Biol 18:1075–1082. doi:10.1038/nsmb.2091

van den Hurk JA et al (2003) Novel types of mutation in the choroideremia (CHM) gene: a full-length L1 insertion and an intronic mutation activating a cryptic exon. Hum Genet 113:268–275. doi:10.1007/s00439-003-0970-0

van Kuilenburg AB et al (2010) Intragenic deletions and a deep intronic mutation affecting pre-mRNA splicing in the dihydropyrimidine dehydrogenase gene as novel mechanisms causing 5-fluorouracil toxicity. Hum Genet 128:529–538. doi:10.1007/s00439-010-0879-3

Varon R et al (2003) Partial deficiency of the C-terminal-domain phosphatase of RNA polymerase II is associated with congenital cataracts facial dysmorphism neuropathy syndrome. Nat Genet 35:185–189. doi:10.1038/ng1243

Vega AI, Perez-Cerda C, Desviat LR, Matthijs G, Ugarte M, Perez B (2009) Functional analysis of three splicing mutations identified in the PMM2 gene: toward a new therapy for congenital disorder of glycosylation type Ia. Hum Mutat 30:795–803. doi:10.1002/humu.20960

Vervoort R, Gitzelmann R, Lissens W, Liebaers I (1998) A mutation (IVS8+0.6kbdelTC) creating a new donor splice site activates a cryptic exon in an Alu-element in intron 8 of the human beta-glucuronidase gene. Hum Genet 103:686–693

Vetrini F et al (2006) Aberrant splicing in the ocular albinism type 1 gene (OA1/GPR143) is corrected in vitro by morpholino antisense oligonucleotides. Hum Mutat 27:420–426. doi:10.1002/humu.20303

Vorechovsky I (2010) Transposable elements in disease-associated cryptic exons. Hum Genet 127:135–154. doi:10.1007/s00439-009-0752-4

Wahl MC, Will CL, Luhrmann R (2009) The spliceosome: design principles of a dynamic RNP machine. Cell 136:701–718. doi:10.1016/j.cell.2009.02.009

Walenkamp MJ et al (2013) Genetic analysis of GHR should contain sequencing of all coding exons and specific intron sequences, and screening for exon deletions. Hormone Res Paediatr 80:406–412. doi:10.1159/000355928

Wang Z, Burge CB (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA 14:802–813. doi:10.1261/rna.876308

Wang GS, Cooper TA (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet 8:749–761. doi:10.1038/nrg2164

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB (2004) Systematic identification and analysis of exonic splicing silencers. Cell 119:831–845. doi:10.1016/j.cell.2004.11.010

Webb TR et al (2012) Deep intronic mutation in OFD1, identified by targeted genomic next-generation sequencing, causes a severe form of X-linked retinitis pigmentosa (RP23). Hum Mol Genet 21:3647–3654. doi:10.1093/hmg/dds194

Will CL, Luhrmann R (2011) Spliceosome structure and function. Cold Spring Harbor Perspect Biol. doi:10.1101/cshperspect.a003707

Will CL, Schneider C, Reed R, Luhrmann R (1999) Identification of both shared and distinct proteins in the major and minor spliceosomes. Science 284:2003–2005

Williams GT, Farzaneh F (2012) Are snoRNAs and snoRNA host genes new players in cancer? Nat Rev Cancer 12:84–88. doi:10.1038/nrc3195

Wilusz JE (2015) Repetitive elements regulate circular RNA biogenesis. Mob Genet Elem 5:1–7. doi:10.1080/2159256X.2015.1045682

Witten JT, Ule J (2011) Understanding splicing regulation through RNA splicing maps. Trends Genet 27:89–97. doi:10.1016/j.tig.2010.12.001

Wong JJ et al (2013) Orchestrated intron retention regulates normal granulocyte differentiation. Cell 154:583–595. doi:10.1016/j.cell.2013.06.052

Wu S, Romfo CM, Nilsen TW, Green MR (1999) Functional recognition of the 3′ splice site AG by the splicing factor U2AF35. Nature 402:832–835. doi:10.1038/45590

Xiong HY et al (2015) RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science 347:1254806. doi:10.1126/science.1254806

Yagi M, Takeshima Y, Wada H, Nakamura H, Matsuo M (2003) Two alternative exons can result from activation of the cryptic splice acceptor site deep within intron 2 of the dystrophin gene in a patient with as yet asymptomatic dystrophinopathy. Hum Genet 112:164–170. doi:10.1007/s00439-002-0854-8

Yap K, Lim ZQ, Khandelia P, Friedman B, Makeyev EV (2012) Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. Genes Dev 26:1209–1223. doi:10.1101/gad.188037.112

Yasmeen S et al (2014) Occipital horn syndrome and classical Menkes Syndrome caused by deep intronic mutations, leading to the activation of ATP7A pseudo-exon. Eur J Hum Genet 22:517–521. doi:10.1038/ejhg.2013.191

Yasuda H, Oh CD, Chen D, de Crombrugghe B, Kim JH (2017) A novel regulatory mechanism of type II collagen expression via a SOX9-dependent enhancer in intron 6. J Biol Chem 292:528–538. doi:10.1074/jbc.M116.758425

Yu Y et al (2008) Dynamic regulation of alternative splicing by silencers that modulate 5′ splice site competition. Cell 135:1224–1236. doi:10.1016/j.cell.2008.10.046

Zamore PD, Patton JG, Green MR (1992) Cloning and domain structure of the mammalian splicing factor U2AF. Nature 355:609–614. doi:10.1038/355609a0

Zhang XH, Chasin LA (2004) Computational definition of sequence motifs governing constitutive exon splicing. Genes Dev 18:1241–1250. doi:10.1101/gad.1195304

Zhang Y et al (2013) Circular intronic long noncoding RNAs. Mol Cell 51:792–806. doi:10.1016/j.molcel.2013.08.017

Zorio DA, Blumenthal T (1999) Both subunits of U2AF recognize the 3′ splice site in *Caenorhabditis elegans*. Nature 402:835–838. doi:10.1038/45597

Aos meus pais. Por regarem os meus sonhos.